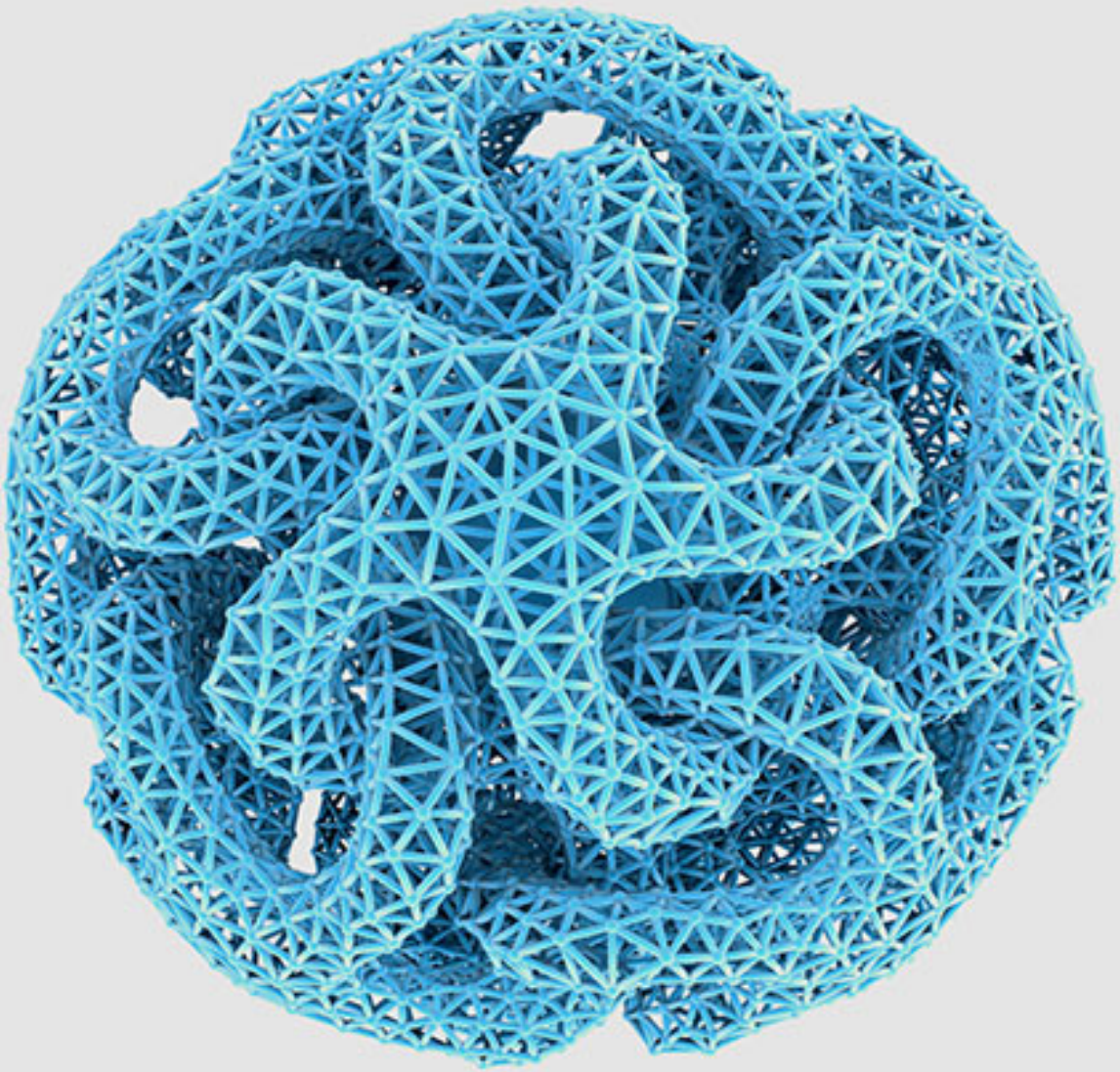
EBOOKS

SCIENTIFIC AMERICAN®

Mathematics IN THE 21st Century



Mathematics in the 21st Century

From the Editors of Scientific American

Cover Image: ALFRED PASIEKA/SCIENCE PHOTO LIBRARY/Getty Images

Letters to the Editor

Scientific American
One New York Plaza
Suite 4500
New York, NY 10004-1562
or editors@sciam.com

Published by Scientific American
www.scientificamerican.com

ISBN: 978-1-948933-12-4

SCIENTIFIC AMERICAN

SCIENTIFIC AMERICAN MIND

SCIENTIFIC AMERICAN®

Mathematics in the 21st Century

From the Editors of Scientific American

Table of Contents

Cracking the Brain's Enigma Code
by Helen Shen

SECTION 3
Mathematics for Understanding the Physical World

3.1
How Einstein Discovered Reality
by Walter Isaacson

3.2
The Strangest Numbers in String Theory
by John C. Baez and John Huerta

3.3
Walls of Water
by Dana Mackenzie

3.4
The Particle Code
by Matthew von Hippel

SECTION 4
Mathematics and Human Culture

4.1
Geometry v. Gerrymandering
by Moon Duchin

4.2
Hacking Passwords with Math
by Jean-Paul Delahaye

4.3
Art by the Numbers
by Stephen Ornes

SECTION 5
The Outer Limits

Mathematics Through the Millenia

In 1960, Nobel Prize-winning physicist Eugene Wigner wrote about the “unreasonable effectiveness of mathematics” to help scientists answer questions in a broad range of disciplines. Sixty years later, that assessment still holds.

Mathematics has changed drastically over the millennia it has been practiced. Until a few hundred years ago, mathematicians usually saw themselves as probing objective truth. Euclid’s *Elements*, the influential 3rd-century BCE geometry text, codified what we now call Euclidean or plane geometry. Its axioms—the starting assumptions on which the rest of the work builds—were chosen to reflect the way objects in the real world behave. Newton developed calculus in order to put his physics on solid ground. The rules of calculus were based on his perceptions of the motion of tangible objects.

Today, however, mathematicians do not necessarily see their axioms as true. They are more interested in the implications given certain axioms, not whether the axioms themselves are valid. Euclidean geometry is what happens with one set of assumptions; hyperbolic geometry, in which “straight” lines bend away from each other, is just as valid if you start with different assumptions. In some ways, mathematics is a collection of thought experiments, chains of reasoning beginning with assumptions that may or may not hold in any given real context. It is an abstract art form with aesthetic considerations of its own.

It is something of a miracle, then, that mathematics is still such a powerful tool in science, where the goal is to test hypotheses about the real world. But the importance of mathematical modeling and statistical analysis in other sciences continues to grow.

In this collection, we present some of the mathematical gems from the pages of Scientific American since the turn of the millennium. We start with

a collection of articles about some of the most important purely mathematical results of the past few decades, the frontiers of the field. In Section 2, we explore the body mathematic: the way mathematical modeling is helping scientists understand biology. Section 3 is about mathematics in service of physics. The two fields have been closely entwined for millennia, and their coevolution continues today. Section 4 is about the role of mathematics in the way human beings relate to each other: politics, art, and of course trying to keep or steal secrets.

Though mathematics has created ever-more-powerful tools for modeling the world and computing with precision, we must also reckon with its limitations. In Section 5, we bring you one article from the deep vaults, a 1956 exploration of Kurt Gödel's groundbreaking incompleteness theorems. He showed that there are questions mathematics will never answer, no matter what axioms we choose and how deeply we commit ourselves to studying them. Since then, mathematicians have continued to probe the limitations of the discipline, seeking to define the very boundaries of what humans can know.

--Evelyn Lamb
Book Editor

SECTION 1

Frontiers of Mathematics

The Whole Universe Catalog

by Stephen Ornes

A seemingly endless variety of food was sprawled over several tables at the home of Judith L. Baxter and her husband, mathematician Stephen D. Smith, in Oak Park, Ill., on a cool Friday evening in September 2011. Canapés, homemade meatballs, cheese plates and grilled shrimp on skewers crowded against pastries, pâtés, olives, salmon with dill sprigs and feta wrapped in eggplant. Dessert choices included—but were not limited to—a lemon mascarpone cake and an African pumpkin cake. The sun set, and champagne flowed, as the 60 guests, about half of them mathematicians, ate and drank and ate some more.

The colossal spread was fitting for a party celebrating a mammoth achievement. Four mathematicians at the dinner—Smith, Michael Aschbacher, Richard Lyons and Ronald Solomon—had just published a book, more than 180 years in the making, that gave a broad overview of the biggest division problem in mathematics history.

Their treatise did not land on any best-seller lists, which was understandable, given its title: *The Classification of Finite Simple Groups*. But for algebraists, the 350-page tome was a milestone. It was the short version, the CliffsNotes, of this universal classification. The full proof reaches some 15,000 pages—some say it is closer to 10,000—that are scattered across hundreds of journal articles by more than 100 authors. The assertion that it supports is known, appropriately, as the Enormous Theorem. (The theorem itself is quite simple. It is the proof that gets gigantic.) The cornucopia at Smith’s house seemed an appropriate way to honor this behemoth. The proof is the largest in the history of mathematics.

And now it is in peril. The 2011 work sketches only an outline of the proof. The unmatched heft of the actual documentation places it on the

teetering edge of human unmanageability. “I don’t know that anyone has read everything,” says Solomon, age 66, who studied the proof his entire career. (He retired from Ohio State University two years ago.) Solomon and the other three mathematicians honored at the party may be the only people alive today who understand the proof, and their advancing years have everyone worried. Smith is 67, Aschbacher is 71 and Lyons is 70. “We’re all getting old now, and we want to get these ideas down before it’s too late,” Smith says. “We could die, or we could retire, or we could forget.”

That loss would be, well, enormous. In a nutshell, the work brings order to group theory, which is the mathematical study of symmetry. Research on symmetry, in turn, is critical to scientific areas such as modern particle physics. The Standard Model—the cornerstone theory that lays out all known particles in existence, found and yet to be found—depends on the tools of symmetry provided by group theory. Big ideas about symmetry at the smallest scales helped physicists figure out the equations used in experiments that would reveal exotic fundamental particles, such as the quarks that combine to make the more familiar protons and neutrons.

Group theory also led physicists to the unsettling idea that mass itself—the amount of matter in an object such as this magazine, you, everything you can hold and see—formed because symmetry broke down at some fundamental level. Moreover, that idea pointed the way to the discovery of the most celebrated particle in recent years, the Higgs boson, which can exist only if symmetry falters at the quantum scale. The notion of the Higgs popped out of group theory in the 1960s but was not discovered until 2012, after experiments at CERN’s Large Hadron Collider near Geneva.

Symmetry is the concept that something can undergo a series of transformations—spinning, folding, reflecting, moving through time—and, at the end of all those changes, appear unchanged. It lurks everywhere in the universe, from the configuration of quarks to the arrangement of galaxies in the cosmos.

The Enormous Theorem demonstrates with mathematical precision that any kind of symmetry can be broken down and grouped into one of four families, according to shared features. For mathematicians devoted to the rigorous study of symmetry, or group theorists, the theorem is an accomplishment no less sweeping, important or fundamental than the

periodic table of the elements was for chemists. In the future, it could lead to other profound discoveries about the fabric of the universe and the nature of reality.

Except, of course, that it is a mess: the equations, corollaries and conjectures of the proof have been tossed amid more than 500 journal articles, some buried in thick volumes, filled with the mixture of Greek, Latin and other characters used in the dense language of mathematics. Add to that chaos the fact that each contributor wrote in his or her idiosyncratic style.

That mess is a problem because without every piece of the proof in position, the entirety trembles. For comparison, imagine the two-million-plus stones of the Great Pyramid of Giza strewn haphazardly across the Sahara, with only a few people who know how they fit together. Without an accessible proof of the Enormous Theorem, future mathematicians would have two perilous choices: simply trust the proof without knowing much about how it works or reinvent the wheel. (No mathematician would ever be comfortable with the first option, and the second option would be nearly impossible.)

The 2011 outline put together by Smith, Solomon, Aschbacher and Lyons was part of an ambitious survival plan to make the theorem accessible to the next generation of mathematicians. “To some extent, most people these days treat the theorem like a black box,” Solomon laments. The bulk of that plan calls for a streamlined proof that brings all the disparate pieces of the theorem together. The plan was conceived more than 30 years ago and is now only half-finished.

If a theorem is important, its proof is doubly so. A proof establishes the honest dependability of a theorem and allows one mathematician to convince another—even when separated by continents or centuries—of the truth of a statement. Then these statements beget new conjectures and proofs, such that the collaborative heart of mathematics stretches back millennia.

Inna Capdeboscq of the University of Warwick in England is one of the few younger researchers to have delved into the theorem. At age 44, soft-spoken and confident, she lights up when she describes the importance of

truly understanding how the Enormous Theorem works. “What is classification? What does it mean to give you a list?” she ponders. “Do we know what every object on this list is? Otherwise, it’s just a bunch of symbols.”

Four Enormous Families

Symmetries can be broken down into basic pieces. Called finite simple groups, they function like elements, coming together in different combinations to form larger, more complicated symmetries.

The Enormous Theorem organizes these groups into four families. Although its proof is huge, the theorem itself is just one sentence that lists all four: “Every finite simple group is cyclic of prime order, an alternating group, a finite simple group of Lie type, or one of the twenty-six sporadic finite simple groups.”

Here is a brief rundown of those families:

Cyclic groups were among the first building blocks to be categorized. Turn a regular pentagon through one fifth of a circle, or 72 degrees, and it looks unchanged. Turn it five times, and you are back at the beginning. Cyclic groups repeat themselves. The cyclic finite simple groups each have a prime number of members. Cyclic groups with more than two even numbers of members can be broken down further, so they are not simple.

Alternating groups come from switching around the members of a set. A full group of symmetries contains all the permutations, or switches. But an alternating group contains only half of them—the ones that have an even number of switches. For example, let us say you had a set of three numbers: 1, 2 and 3. There are six different ways to write that set: (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), and (3, 2, 1). The alternating group contains three of those. In terms of symmetry, each of these arrangements might correspond to a sequence of symmetries (that is, turn the cube up, then on its side, and so on).

Lie-type groups, named for 19th-century mathematician Sophus Lie, start to get more complicated. They are related to things called infinite Lie groups. The infinite groups include the rotations of a space itself that do not change the volume. For example, there are infinitely many ways to spin a doughnut without changing the doughnut itself. The finite analogues of these infinite groups are the Lie-type groups—in other words, the doughnut in a Lie-type group permits only a finite number of rotations. Most finite simple groups fall into this family. Neither infinite Lie groups nor Lie-type groups are limited to our pedestrian three dimensions. Ready to talk about the symmetries that arise in 15-dimensional space? Then look to these groups.

Sporadic groups make up the family of rogues. They include 26 outliers that do not line up neatly in the other families. (Imagine if the periodic table of elements had a column for “miscreants.”) The largest of these sporadic groups, called the Monster, has more than 10^{53} elements and can be faithfully represented in 196,883 dimensions. It is baffling and bizarre, and no one really knows what it means but it is tantalizing to think about. “I have a sneaking hope, a hope unsupported by any facts or any evidence,” physicist Freeman Dyson wrote in 1983, “that sometime in the twenty-first century physicists will stumble upon the Monster group, built in some unsuspected way into the structure of the universe.”

Reality's Deepest Secrets

Mathematicians first began dreaming of the proof at least as early as the 1890s, as a new field called group theory took hold. In math, the word “group” refers to a set of objects connected to one another by some mathematical operation. If you apply that operation to any member of the group, the result is yet another member.

Symmetries, or movements that do not change the look of an object, fit this bill. Consider, as an example, that you have a cube with every side painted the same color. Spin the cube 90 degrees—or 180 or 270—and the cube will look exactly as it did when you started. Flip it over, top to bottom, and it will appear unchanged. Leave the room and let a friend spin or flip the cube—or execute some combination of spins and flips—and when you return, you will not know what he or she has done. In all, there are 24 distinct rotations that leave a cube appearing unchanged. Those 24 rotations make a finite group.

Simple finite groups are analogous to atoms. They are the basic units of construction for other, larger things. Simple finite groups combine to form larger, more complicated finite groups. The Enormous Theorem organizes these groups the way the periodic table organizes the elements. It says that every simple finite group belongs to one of three families—or to a fourth family of wild outliers. The largest of these rogues, called the Monster, has more than 1053 elements and exists in 196,883 dimensions. (There is even a whole field of investigation called monsterology in which researchers search for signs of the beast in other areas of math and science.) The first finite simple groups were identified by 1830, and by the 1890s mathematicians had made new inroads into finding more of those building blocks. Theorists also began to suspect the groups could all be put together in a big list.

Mathematicians in the early 20th century laid the foundation for the Enormous Theorem, but the guts of the proof did not materialize until midcentury. Between 1950 and 1980—a period which mathematician Daniel Gorenstein of Rutgers University called the “Thirty Years’ War”—heavyweights pushed the field of group theory further than ever before, finding finite simple groups and grouping them together into families. These mathematicians wielded 200-page manuscripts like algebraic machetes, cutting away abstract weeds to reveal the deepest foundations of

symmetry. (Freeman Dyson of the Institute for Advanced Study in Princeton, N.J., referred to the onslaught of discovery of strange, beautiful groups as a “magnificent zoo.”)

Those were heady times: Richard Foote, then a graduate student at the University of Cambridge and now a professor at the University of Vermont, once sat in a dank office and witnessed two famous theorists—John Thompson, now at the University of Florida, and John Conway, now at Princeton University—hashing out the details of a particularly unwieldy group. “It was amazing, like two Titans with lightning going between their brains,” Foote says. “They never seemed to be at a loss for some absolutely wonderful and totally off-the-wall techniques for doing something. It was breathtaking.”

It was during these decades that two of the proof’s biggest milestones occurred. In 1963 a theorem by mathematicians Walter Feit and John Thompson laid out a recipe for finding more simple finite groups. After that breakthrough, in 1972 Gorenstein laid out a 16-step plan for proving the Enormous Theorem—a project that would, once and for all, put all the finite simple groups in their place. It involved bringing together all the known finite simple groups, finding the missing ones, putting all the pieces into appropriate categories and proving there could not be any others. It was big, ambitious, unruly and, some said, implausible.

The Man with the Plan

Yet Gorenstein was a charismatic algebraist, and his vision energized a new group of mathematicians—with ambitions neither simple nor finite—who were eager to make their mark. “He was a larger than life personality,” says Lyons, who is at Rutgers. “He was tremendously aggressive in the way he conceived of problems and conceived of solutions. And he was very persuasive in convincing other people to help him.”

Solomon, who describes his first encounter with group theory as “love at first sight,” met Gorenstein in 1970. The National Science Foundation was hosting a summer institute on group theory at Bowdoin College, and every week mathematical celebrities were invited to the campus to give a lecture. Solomon, who was then a graduate student, remembers Gorenstein’s visit

vividly. The mathematical celebrity, just arrived from his summer home on Martha's Vineyard, was electrifying in both appearance and message.

"I'd never seen a mathematician in hot-pink pants before," Solomon recalls.

In 1972, Solomon says, most mathematicians thought that the proof would not be done by the end of the 20th century. But within four years the end was in sight. Gorenstein largely credited the inspired methods and feverish pace of Aschbacher, who is a professor at the California Institute of Technology, for hastening the proof's completion.

One reason the proof is so huge is that it stipulates that its list of finite simple groups is complete. That means the list includes every building block, and there are not any more. Oftentimes proving something does not exist—such as proving there cannot be any more groups—is more work than proving it does.

In 1981 Gorenstein declared the first version of the proof finished, but his celebration was premature. A problem emerged with a particularly thorny 800-page chunk, and it took some debate to resolve it successfully. Mathematicians occasionally claimed to find other flaws in the proof or to have found new groups that broke the rules. To date, those claims have failed to topple the proof, and Solomon says he is fairly confident that it will stand.

Gorenstein soon saw the theorem's documentation for the sprawling, disorganized tangle that it had become. It was the product of a haphazard evolution. So he persuaded Lyons—and in 1982 the two of them ambushed Solomon—to help forge a revision, a more accessible and organized presentation, which would become the so-called second-generation proof. Their goals were to lay out its logic and keep future generations from having to reinvent the arguments, Lyons says. In addition, the effort would whittle the proof's 15,000 pages down, reducing it to a mere 3,000 or 4,000.

Gorenstein envisioned a series of books that would neatly collect all the disparate pieces and streamline the logic to iron over idiosyncrasies and eliminate redundancies. In the 1980s the proof was inaccessible to all but the seasoned veterans of its forging. Mathematicians had labored on it for decades, after all, and wanted to be able to share their work with future

generations. A second-generation proof would give Gorenstein a way to assuage his worries that their efforts would be lost amid heavy books in dusty libraries.

Gorenstein did not live to see the last piece put in place, much less raise a glass at the Smith and Baxter house. He died of lung cancer on Martha's Vineyard in 1992. "He never stopped working," Lyons recalls. "We had three conversations the day before he died, all about the proof. There were no good-byes or anything; it was all business."

Proving It Again

The first volume of the second-generation proof appeared in 1994. It was more expository than a standard math text and included only two of 30 proposed sections that could entirely span the Enormous Theorem. The second volume was published in 1996, and subsequent ones have continued to the present—the sixth appeared in 2005.

Foote says the second-generation pieces fit together better than the original chunks. "The parts that have appeared are more coherently written and much better organized," he says. "From a historical perspective, it's important to have the proof in one place. Otherwise, it becomes sort of folklore, in a sense. Even if you believe it's been done, it becomes impossible to check."

Solomon and Lyons are finishing the seventh book this summer, and a small band of mathematicians have already made inroads into the eighth and ninth. Solomon estimates that the streamlined proof will eventually take up 10 or 11 volumes, which means that just more than half of the revised proof has been published.

Solomon notes that the 10 or 11 volumes *still* will not entirely cover the second-generation proof. Even the new, streamlined version includes references to supplementary volumes and previous theorems, proved elsewhere. In some ways, that reach speaks to the cumulative nature of mathematics: every proof is a product not only of its time but of all the thousands of years of thought that came before.

In a 2005 article in the *Notices of the American Mathematical Society*, mathematician E. Brian Davies of King's College London pointed out that

the “proof has never been written down in its entirety, may never be written down, and as presently envisaged would not be comprehensible to any single individual.” His article brought up the uncomfortable idea that some mathematical efforts may be too complex to be understood by mere mortals. Davies’s words drove Smith and his three co-authors to put together the comparatively concise book that was celebrated at the party in Oak Park.

The Enormous Theorem’s proof may be beyond the scope of most mathematicians—to say nothing of curious amateurs—but its organizing principle provides a valuable tool for the future. Mathematicians have a long-standing habit of proving abstract truths decades, if not centuries, before they become useful outside the field.

“One thing that makes the future exciting is that it is difficult to predict,” Solomon observes. “Geniuses come along with ideas that nobody of our generation has had. There is this temptation, this wish and dream, that there is some deeper understanding still out there.”

The Next Generation

These decades of deep thinking did not only move the proof forward; they built a community. Judith Baxter—who trained as a mathematician—says group theorists form an unusually social group. “The people in group theory are often lifelong friends,” she observes. “You see them at meetings, travel with them, go to parties with them, and it is really is a wonderful community.”

Not surprisingly, these mathematicians who lived through the excitement of finishing the first iteration of the proof are eager to preserve its ideas. Accordingly, Solomon and Lyons have recruited other mathematicians to help them finish the new version and preserve it for the future. That is not easy: many younger mathematicians see the proof as something that has already been done, and they are eager for something different.

In addition, working on rewriting a proof that has already been established takes a kind of reckless enthusiasm for group theory. Solomon found a familiar devotee to the field in Capdeboscq, one of a handful of younger mathematicians carrying the torch for the completion of the second-generation proof. She became enamored of group theory after taking a class from Solomon.

“To my surprise, I remember reading and doing the exercises and thinking that I loved it. It was beautiful,” Capdeboscq says. She got “hooked” on working on the second-generation proof after Solomon asked for her help in figuring out some of the missing pieces that would eventually become part of the sixth volume. Streamlining the proof, she says, lets mathematicians look for more straightforward approaches to difficult problems.

Capdeboscq likens the effort to refining a rough draft. Gorenstein, Lyons and Solomon laid out the plan, but she says it is her job, and the job of a few other youngsters, to see all the pieces fall into place: “We have the road map, and if we follow it, at the end the proof should come out."

-- Originally published: Scientific American 313(1); 68-75 (July 2015).

The Shapes of Space

by Graham P. Collins

Stand up and look around. Walk in a circle. Jump in the air. Wave your arms. You are a collection of particles moving about within a small region of a 3-manifold—a three-dimensional space—that extends in all directions for many billions of light-years.

Manifolds are mathematical constructs. The triumph of physics since the time of Galileo and Kepler has been the successful description of reality by mathematics of one flavor or another, such as the mathematics of manifolds. According to physics, everything that happens takes place against the backdrop of three-dimensional space (leaving aside the speculations of string theorists that there are tiny dimensions in addition to the three that are manifest). Three dimensions means that three numbers are needed to specify the location of a particle. Near Earth, for instance, the three numbers could be latitude, longitude and altitude.

According to Newtonian physics and traditional quantum physics, the three-dimensional space where everything happens is fixed and immutable. Einstein's theory of general relativity, in contrast, makes space an active player: the distance from one point to another is influenced by how much matter and energy are nearby and by any gravitational waves that may be passing by. But whether we are dealing with Newtonian or Einsteinian physics and whether space is infinite or finite, space is represented by a 3-manifold. Understanding the properties of 3-manifolds is therefore essential for fully comprehending the foundations on which almost all of physics—and all other sciences—are built. (The 4-manifolds are also important: space and time together form a 4-manifold.)

Mathematicians know a lot about 3-manifolds, yet some of the most basic questions have proved to be the hardest. The branch of mathematics that

studies manifolds is topology. Among the fundamental questions topologists can ask about 3-manifolds are: What is the simplest type of 3-manifold, the one with the least complicated structure? Does it have many cousins that are equally simple, or is it unique? What kinds of 3-manifolds are there?

The answer to the first of those questions has long been known: a space called the 3-sphere is the simplest compact 3-manifold. (Noncompact manifolds can be thought of as being infinite or having an edge. Hereafter I consider only compact manifolds.) The other two questions have been up for grabs for a century but may have been answered in 2002 by Grigori ("Grisha") Perelman, a Russian mathematician who has most probably proved a theorem known as the Poincaré conjecture.

First postulated by French mathematician Henri Poincaré exactly 100 years ago, the conjecture holds that the 3-sphere is unique among 3-manifolds; no other 3-manifold shares the properties that make it so simple. The 3-manifolds that are more complicated than the 3- sphere have boundaries that you can run up against like a brick wall, or multiple connections from one region to another, like a path through the woods that splits and later rejoins. The Poincaré conjecture states that the 3-sphere is the only compact 3-manifold that lacks all those complications. Any three-dimensional object that shares those properties with the sphere can therefore be morphed into the same shape as a 3-sphere; so far as topologists are concerned, the object is just another copy of the 3-sphere. Perelman's proof also answers the third of our questions: it completes work that classifies all the types of 3-manifolds that exist.

It takes some mental gymnastics to imagine what a 3-sphere is like—it is not simply a sphere in the everyday sense of the word. But it has many properties in common with the 2-sphere, which we are all familiar with: If you take a spherical balloon, the rubber of the balloon forms a 2-sphere. The 2-sphere is two-dimensional because only two coordinates—latitude and longitude—are needed to specify a point on it. Also, if you take a very small disk of the balloon and examine it with a magnifying glass, the disk looks a lot like one cut from a flat two-dimensional plane of rubber. It just has a slight curvature. To a tiny insect crawling on the balloon, it would seem like a flat plane. Yet if the insect traveled far enough in what would

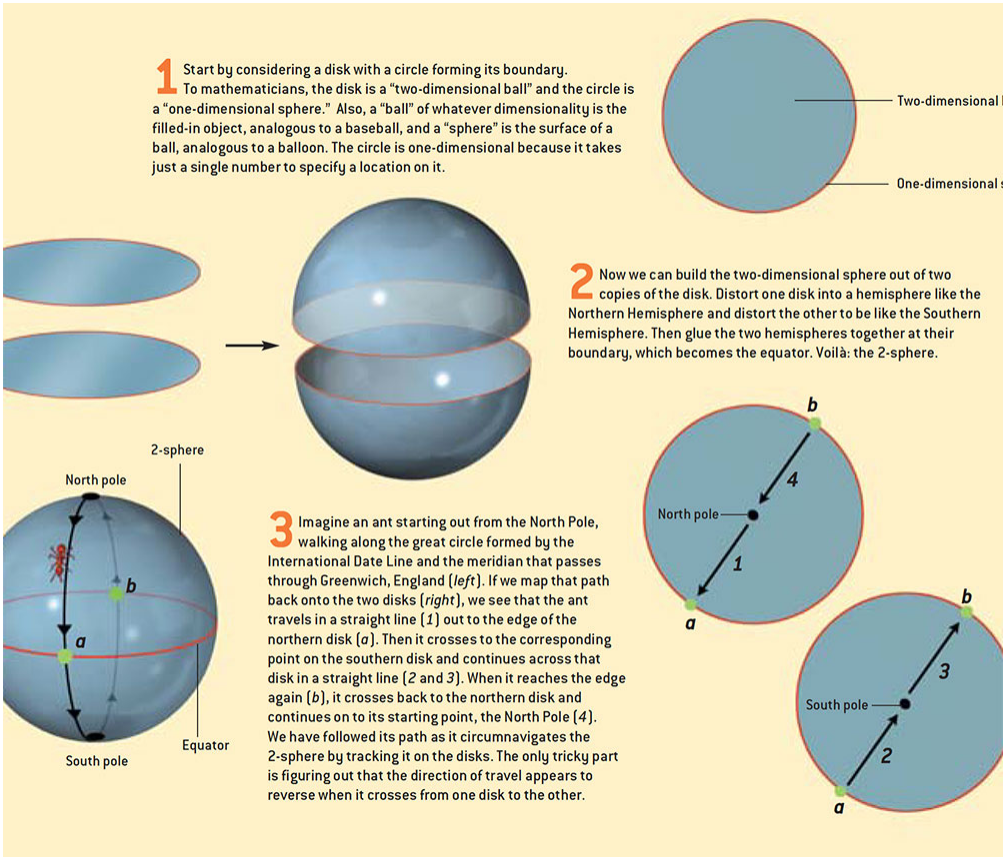
seem to it to be a straight line, eventually it would arrive back at its starting point.

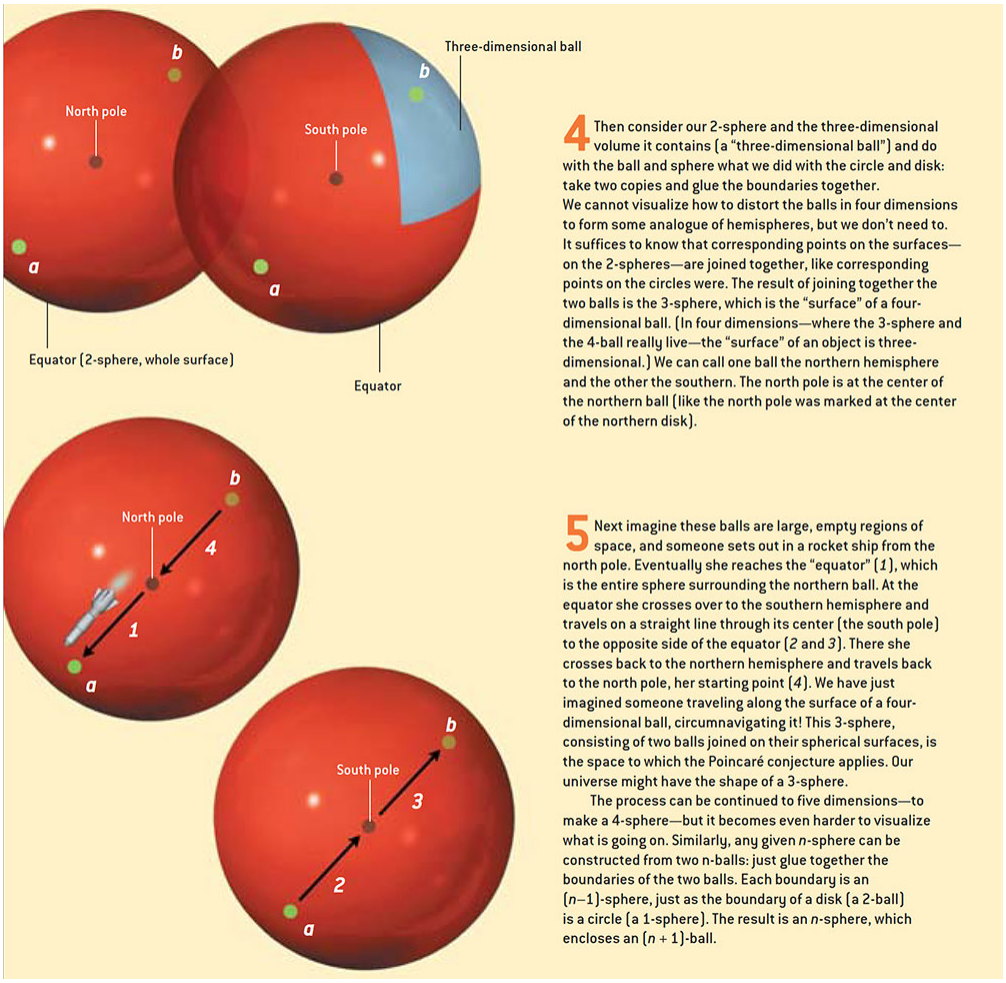
Similarly, a gnat in a 3-sphere—or a person in one as big as our universe!—perceives itself to be in "ordinary" three-dimensional space. But if it flies far enough in a straight line in any direction, eventually it will circumnavigate the 3-sphere and find itself back where it started, just like the insect on the balloon or someone taking a trip around the world.

Spheres exist for dimensions other than three as well. The 1-sphere is also familiar to you: it is just a circle (the rim of a disk, not the disk itself). The *n*-dimensional sphere is called an *n*-sphere.

Multidimensional Music of Spheres

The 3-sphere at the heart of Poincaré's conjecture takes a bit of effort to visualize. Mathematicians who prove theorems about higher-dimensional spaces do not have to visualize them. They make do with abstract properties, guided by intuitive notions based on analogies to lower dimensions (but being careful not to take the analogies literally). Others, too, can form an idea of what higher-dimensional objects are like by working up from familiar lower-dimensional examples. The 3-sphere is a case in point.





Proving Conjectures

After Poincaré proposed his conjecture about the 3-sphere, half a century went by before any real progress was made in proving it. In the 1960s mathematicians proved analogues of the conjecture for spheres of five dimensions or more. In each case, the *n*-sphere is the unique, simplest manifold of that dimensionality. Paradoxically, this result was easier to prove for higher-dimensional spheres than for those of four or three dimensions. The proof for the particularly difficult case of four dimensions came in 1982. Only the original three-dimensional case involving Poincaré's 3-sphere remained open.

A major step in closing the three-dimensional problem came in November 2002, when Perelman, a mathematician at the Steklov Institute of Mathematics at St. Petersburg, posted a paper on the www.arxiv.org Web server that is widely used by physicists and mathematicians as a

clearinghouse of new research. The paper did not mention the Poincaré conjecture by name, but topology experts who looked at it immediately realized the paper's relevance to that theorem. Perelman followed up with a second paper in March 2003, and from April to May that year he visited the U.S. to give a series of seminars on his results at the Massachusetts Institute of Technology and Stony Brook University. Teams of mathematicians at nearly a dozen leading institutes began poring over his papers, verifying their every detail and looking for errors.

At Stony Brook, Perelman gave two weeks of formal and informal lectures, talking from three to six hours a day. "He answered every question that arose, and he was very clear," says mathematician Michael Anderson of Stony Brook. "No one has yet raised any serious doubts." One more comparatively minor step has to be proved to complete the result, Anderson says, "but there are no real doubts about the validity of this final piece." The first paper contains the fundamental ideas and is pretty well accepted as being verified. The second paper contains applications and more technical arguments; its verification has not reached the level of confidence achieved for the first paper.

The Poincaré conjecture has a $1-million reward on offer for its proof: it is one of seven such "Millennium Problems" singled out in 2000 by the Clay Mathematics Institute in Cambridge, Mass. Perelman's proof has to be published and withstand two years of scrutiny before he becomes eligible for the prize. (The institute might well decide that its posting on the Web server qualifies as "published" because the result is undergoing as rigorous a peer review as any paper gets.)

Perelman's work extends and completes a program of research that Richard S. Hamilton of Columbia University explored in the 1990s. The Clay Institute recognized Hamilton's work with a research award in late 2003. Perelman's calculations and analysis blow away several roadblocks that Hamilton ran into and could not overcome.

If, as everyone expects, Perelman's proof is correct, it actually completes a much larger body of work than the Poincaré conjecture. Launched by William P. Thurston—now at Cornell University—the Thurston geometrization conjecture provides a full classification of all possible 3-manifolds. The 3-sphere, unique in its sublime simplicity, anchors the

foundation of this magnificent classification. Had the Poincaré conjecture been false—that is, if there were many spaces as “simple” as a sphere—the classification of 3-manifolds would have exploded into something infinitely more complicated than that proposed by Thurston. Instead, with Perelman’s and Thurston’s results, we now have a complete catalogue of all the possible shapes that a three-dimensional space can take on—all the shapes allowed by mathematics that our universe (considering just space and not time) could have.

Rubber Doughnuts

To understand the Poincaré conjecture and Perelman’s proof in greater depth, you have to know something about topology. In that branch of mathematics the exact shape of an object is irrelevant, as if it were made of play dough that you could stretch, compress and bend to any extent. But why should we care about objects or spaces made of imaginary play dough? The reason relates to the fact that the exact shape of an object—the distance from one point on it to another—is a level of structure, which is called the geometry of the object. By considering a play-dough object, topologists discover which properties of the object are so fundamental that they exist independently of its geometric structure. Studying topology is like discovering which properties humans have in general by considering the properties of a “playdough person” who can be morphed into any particular human being.

If you have read any popular account of topology, you have probably encountered the hoary old truism that a cup and a doughnut are indistinguishable to a topologist. (The saying refers to a ring-shaped doughnut, not the solid, jam-filled kind.) The point is that you can morph the play-dough cup into a doughnut shape simply by smushing the clay around, without having to cut out any holes or glue any patches together. A ball, on the other hand, can be turned into a doughnut only by either boring a hole through its middle or stretching it into a cylinder and gluing the ends together. Because such cutting or gluing is needed, a ball is not the same as a doughnut to a topologist.

What interests topologists most are the surfaces of the ball and the doughnut, so instead of imagining a solid we should imagine a balloon in both cases. The topologies are still distinct—the spherical balloon cannot be

morphed into the ring-shaped balloon, which is called a torus. Topologically, then, a sphere and a torus are distinct entities. Early topologists set out to discover how many other topologically distinct entities exist and how they could be characterized. For two-dimensional objects, which are also called surfaces, the answer is neat and tidy: it all boils down to how many “handles” a surface has.

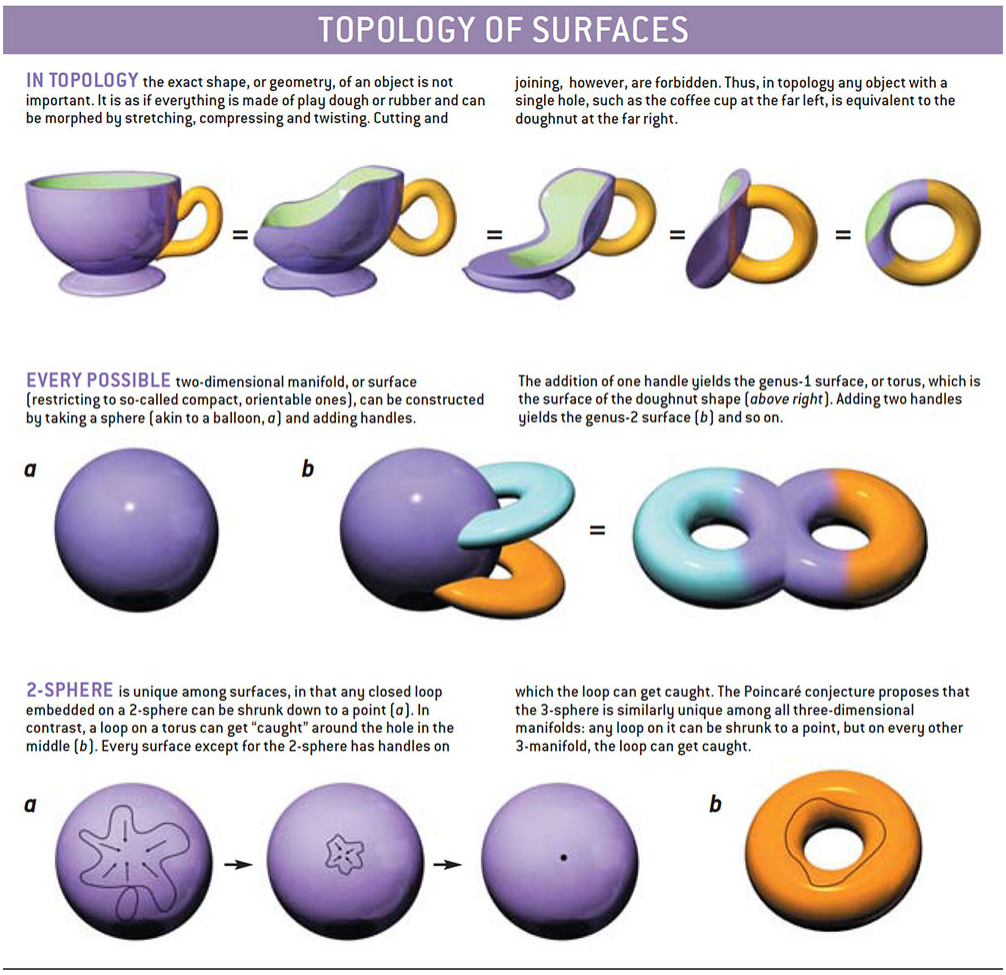
By the end of the 19th century, mathematicians understood how to classify surfaces. Out of all the surfaces, the sphere, they knew, had a unique simplicity. Naturally they started wondering about three-dimensional manifolds. To start with, was the 3-sphere unique in its simplicity, analogous to the 2-sphere? The century-long history that follows from that elementary question is littered with false steps and false proofs.

Henri Poincaré tackled this question head-on. He was one of the two foremost mathematicians who were active at the turn of the 20th century (the other being David Hilbert). Poincaré has been called the last universalist—he was at ease in all branches of mathematics, both pure and applied. In addition to advancing numerous areas of mathematics, he contributed to the theories of celestial mechanics and electromagnetism as well as to the philosophy of science (about which he wrote several widely read popular books).

Poincaré largely created the branch of mathematics called algebraic topology. Around 1900, using techniques from that field, he formulated a measure of an object’s topology, called homotopy. To determine a manifold’s homotopy, imagine that you embed a closed loop in the manifold. The loop can be wound around the manifold in any possible fashion. We then ask, Can the loop always be shrunk down to a point, just by moving it around, without ever lifting a piece of it out of the manifold? On a torus the answer is no. If the loop runs around the circumference of the torus, it cannot be shrunk to a point—it gets caught on the inner ring of the doughnut. Homotopy is a measure of all the different ways a loop can get caught.

On an *n*-sphere, no matter how convoluted a path the loop takes, it can always be untangled and shrunk to a point. (The loop is allowed to pass through itself during these manipulations.) Poincaré speculated that the only 3-manifold on which every possible loop can be shrunk to a point was the

3-sphere itself, but he could not prove it. In due course this proposal became known as the Poincaré conjecture. Over the decades, many people have announced proofs of the conjecture, only to be proved wrong. (For clarity, here and throughout I ignore two complications: so-called nonorientable manifolds and manifolds with edges. For example, the Möbius band, a ribbon that is twisted and joined in a loop, is nonorientable. A sphere with a disk cut out from it has an edge. The Möbius band also has an edge.)



Geometrization

Perelman's proof is the first to withstand close scrutiny. His approach to analyzing three-dimensional manifolds is related to a procedure called geometrization. Geometry relates to the actual shape of an object or manifold: for geometry, an object is made not of play dough but of ceramic.

A cup, for example, has a different geometry than a doughnut; its surface curves in different ways. It is said that the cup and the doughnut are two examples of a topological torus (provided the cup has one handle) to which different geometries have been assigned.

To gain a sense of how geometrization served to help Perelman, consider how geometry can be used to classify 2-manifolds, or surfaces. Each topological surface is assigned a special, unique geometry: the one for which the curvature of the surface is spread completely evenly about the manifold. For the sphere, that unique geometry is the perfectly spherical sphere. An eggshell shape is another possible geometry for a topological sphere, but it does not have curvature evenly spread throughout: the small end of the egg is more curved than the big end.

The 2-manifolds form three geometric types. The sphere has what is called positive curvature, the shape of a hilltop. The geometrized torus is flat; it has zero curvature, like a plain. All the other manifolds, with two or more handles, have negative curvature. Negative curvature is like the shape of a mountain pass or a saddle: going from front to back, a saddle curves up; from left to right, it curves down. Poincaré (who else?), along with Paul Koebe and Felix Klein (after whom the Klein bottle is named), contributed to this geometric classification, or geometrization, of 2-manifolds.

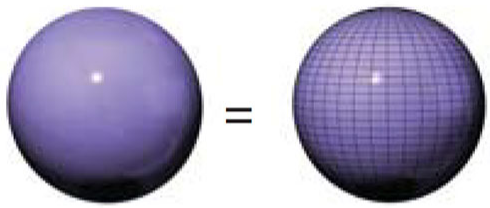
It is natural to try to apply similar methods to 3-manifolds. Is it possible to find unique geometries for each topological 3-manifold, for which curvature is spread evenly throughout the manifold?

It turns out that 3-manifolds are far more complicated than 2-manifolds. Most 3-manifolds cannot be assigned a uniform geometry. Instead they have to be cut up into pieces, each piece having a different canonical geometry. Furthermore, instead of three basic geometries, as with 2-manifolds, the 3-manifold pieces can take any of eight canonical geometries. The cutting up of each 3-manifold is somewhat analogous to the factorization of a number into a unique product of prime factors.

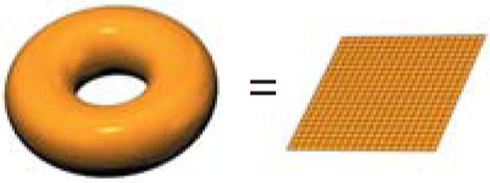
GEOMETRIZATION

2-MANIFOLDS can be classified by "uniformizing" or "geometrizing" them, which means assigning them a specific geometry, or rigid shape. In particular, each can be morphed into a shape that has its curvature evenly distributed. The sphere (*a*) is the unique shape having constant positive curvature, meaning at every point it is curved like a hilltop. The torus (*b*) can be made flat—that is, with zero curvature throughout. To see this, imagine cutting the torus and straightening it out to form a cylinder. Then cut along the cylinder and unroll it to form a flat rectangular plane. The torus has thus been mapped to a flat plane. Surfaces of genus-2 and higher (*c*) can be given constant negative curvature, with other details depending on how many handles are present. Here the constant negative curvature surface is represented by the saddle shape.

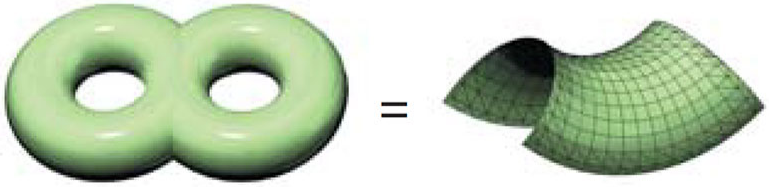
a



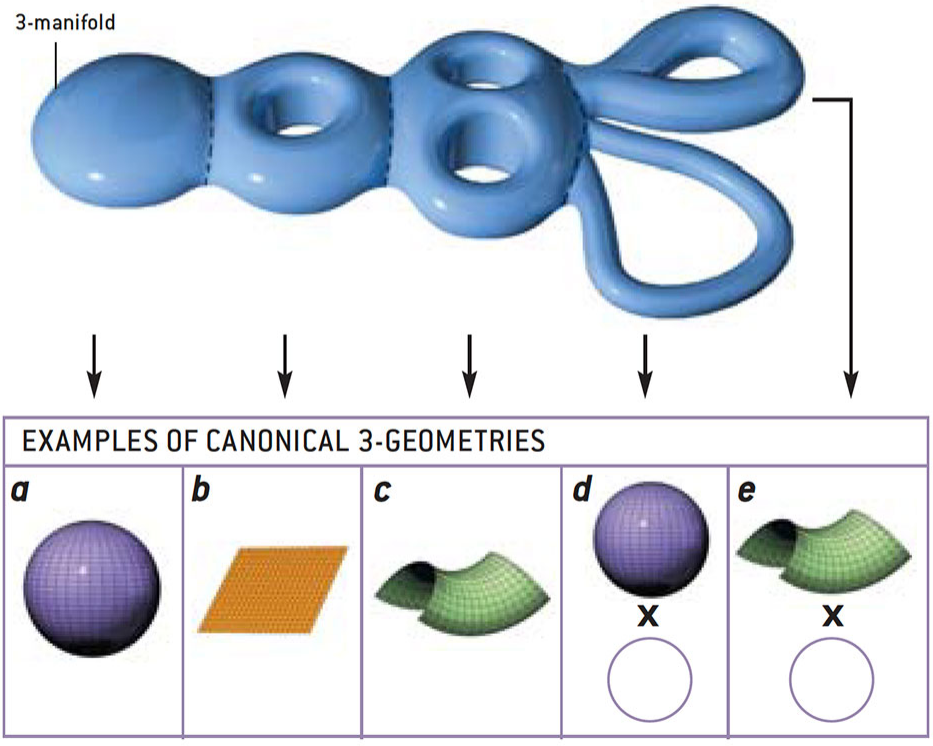
b



c



CLASSIFICATION OF 3-MANIFOLDS, which is similar to that of 2-manifolds but far more complicated, is completed by Perelman's work. In general, a 3-manifold has to be divided into pieces, each of which can be morphed into one of eight different canonical three-dimensional geometries. The blue-colored example below (*drawn in cartoon form as a 2-manifold*) consists of equivalents to five of them: constant positive (*a*), zero (*b*) and constant negative (*c*) curvature 3-geometries, as well as the "product" of the 2-sphere and a circle (*d*) and of the negative curvature surface and a circle (*e*).



This classification scheme was first conjectured by Thurston in the late 1970s. He and his colleagues proved large swaths of the conjecture, but crucial points that the entire system depended on remained beyond their grasp, including the part that embodied the Poincaré conjecture. Was the 3-sphere unique? An answer to that question and completion of the Thurston program have come only with Perelman's papers.

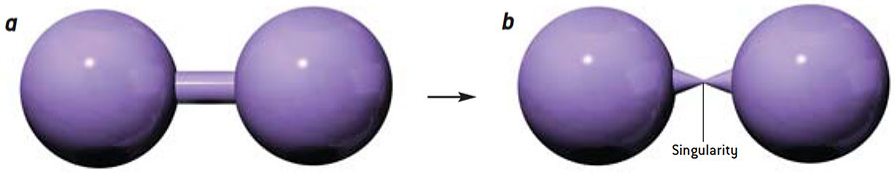
How might we try to geometrize a manifold—that is, give it a uniform curvature throughout? One way is to start with some arbitrary geometry, perhaps like an eggshell shape with various lumps and indentations, and then smooth out all the irregularities. Hamilton began such a program of analysis for 3-manifolds in the early 1990s, using an equation called the Ricci flow (named after mathematician Gregorio Ricci-Curbastro), which has some similarities to the equation that governs the flow of heat. In a

body with hot and cold spots, heat naturally flows from the warmer regions to the cooler ones, until the temperature is uniform everywhere. The Ricci flow equation has a similar effect on curvature, morphing a manifold to even out all the bumps and hollows. If you began with an egg, it would gradually become perfectly spherical.

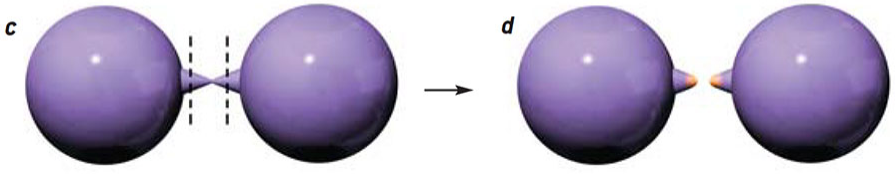
Hamilton's analysis ran into a stumbling block: in certain situations the Ricci flow would cause a manifold to pinch down to a point. (This is one way that the Ricci flow differs from heat flow. The places that are pinched are like points that manage to acquire infinite temperature.) One example was when the manifold had a dumbbell shape, like two spheres connected by a thin neck. The spheres would grow, in effect drawing material from the neck, which would taper to a point in the middle. Another possible example arose when a thin rod stuck out from the manifold; the Ricci flow might produce a trouble called a cigar singularity. When a manifold is pinched in this way, it is called singular—it is no longer a true three-dimensional manifold. In a true three-dimensional manifold, a small region around any point looks like a small region of ordinary three-dimensional space, but this property fails at pinched points. A way around this stumbling block had to wait for Perelman.

DEALING WITH SINGULARITIES

EFFORTS TO APPLY an equation called the Ricci flow to prove Poincaré's conjecture and to geometrize 3-manifolds hit a roadblock before Perelman came along. The Ricci flow, which gradually changes the shape of a 3-manifold, occasionally runs into trouble called singularities. One instance is when part of the manifold has a dumbbell shape—two spheres joined by a tube (*a*). The tube can become pinched down to a point, spoiling the manifold's properties (*b*). Another singularity, called the cigar singularity, was also believed to be possible.



"SURGERY" CAN HANDLE the singularities that arise in the Ricci flow, as shown by Perelman's work. When a region of the manifold starts to pinch down, a small region on each side of the pinch can be cut out (*c*). The cuts are then capped by small spheres, and the Ricci flow is continued (*d*). This process may have to be repeated several times if other regions subsequently pinch down, but Perelman showed that the process terminates eventually. He also showed that cigar singularities never arise.



Perelman came to the U.S. as a postdoctoral student in 1992, spending semesters at New York University and Stony Brook, followed by two years

at the University of California at Berkeley. He quickly made a name for himself as a brilliant young star, proving many important and deep results in a particular branch of geometry. He was awarded a prize from the European Mathematical Society (which he declined) and received a prestigious invitation to address the International Congress of Mathematicians (which he accepted). In spring 1995 he was offered positions at a number of outstanding mathematics departments, but he turned them all down to return to his home in St. Petersburg. “Culturally, he is very Russian,” commented one American colleague. “He’s very unmaterialistic.”

Back in St. Petersburg, he essentially disappeared from mathematicians’ radar screens. The only signs of activity, after many years, were rare occasions when he e-mailed former colleagues, for example, to point out errors in papers they had posted on the Internet. E-mails inquiring about his pursuits went unanswered.

Finally, in late 2002 several people received an e-mail from him alerting them to the paper he had posted on the mathematics server—just a characteristically brief note saying they might find it of interest. That understatement heralded the first stage of his attack on the Poincaré conjecture. In the preprint, along with his Steklov Institute affiliation, Perelman acknowledged support in the form of money he had saved from his U.S. postdoctoral positions.

In his paper, Perelman added a new term to the Ricci flow equation. The modified equation did not eliminate the troubles with singularities, but it enabled Perelman to carry the analysis much further. With the dumbbell singularities he showed that “surgery” could be performed: Snip the thin tube on each side of the incipient pinch and seal off the open tube on each dumbbell ball with a spherical cap. Then the Ricci flow could be continued with the surgically altered manifold until the next pinch, for which the same procedure could be applied. He also showed that cigar singularities could not occur. In this way, any 3-manifold could be reduced to a collection of pieces, each having a uniform geometry.

When the Ricci flow and the surgery are applied to all possible 3-manifolds, any manifold that is as “simple” as a 3-sphere (technically, that has the same homotopy as a 3-sphere) necessarily ends up with the same

uniform geometry as a 3-sphere. That result means that topologically, the manifold in question *is* a 3-sphere. Rephrasing that, the 3-sphere is unique.

Beyond proving Poincaré's conjecture, Perelman's research is important for the innovative techniques of analysis it has introduced. Already mathematicians are posting papers that build on his work or apply his techniques to other problems. In addition, the mathematics has curious connections to physics. The Ricci flow used by Hamilton and Perelman is related to something called the renormalization group, which specifies how interactions change in strength depending on the energy of a collision. For example, at low energies the electromagnetic interaction has a strength characterized by the number 0.0073 (about $1/137$). If two electrons collide head-on at nearly the speed of light, however, the strength is closer to 0.0078.

Increasing the collision energy is equivalent to studying the force at a shorter distance scale. The renormalization group is therefore like a microscope with a magnification that can be turned up or down to examine a process at finer or coarser detail. Similarly, the Ricci flow is like a microscope for looking at a manifold at a chosen magnification. Bumps and hollows visible at one magnification disappear at another. Physicists expect that on a scale of about 10^{-35} meter, or the Planck length, the space in which we live will look very different—like a "foam" with many loops and handles and other topological structures. The mathematics that describes how the physical forces change is very similar to that which describes geometrization of a manifold.

Another connection to physics is that the equations of general relativity, which describe the workings of gravity and the large-scale structure of the universe, are closely related to the Ricci flow equation. Furthermore, the term that Perelman added to the basic flow used by Hamilton arises in string theory, which is a quantum theory of gravity. It remains to be seen if his techniques will reveal interesting new information about general relativity or string theory. If that is the case, Perelman will have taught us not only about the shapes of abstract 3-spaces but also about the shape of the particular space in which we live.

--Originally published: Scientific American 291(1); 94-103 (July 2004).

Sphere Packing in the nth Dimension

by Evelyn Lamb

In 1611 German mathematician Johannes Kepler made a conjecture about the densest way to stack oranges or other spheres with a minimum of space between them. It seemed nothing could beat the standard produce stand configuration, but he could not prove it for sure. Four hundred years later University of Pittsburgh mathematician Thomas Hales finally showed that the grocers were right all along. But the question of how to pack spheres most tightly is not confined to our measly three dimensions—mathematicians can also imagine the problem in hypothetical spaces of any number of dimensions.

In March Ukrainian mathematician Maryna Viazovska, a postdoctoral researcher at the Berlin Mathematical School and at Humboldt University of Berlin, solved the sphere-packing problem in eight dimensions. The next week she and several co-authors extended her techniques to 24 dimensions. The solution of the problem in the seemingly arbitrary dimensions of eight and 24 highlights the fundamental weirdness of sphere packing, which has now been solved only in dimensions one, two, three, eight and 24. The breakthrough has given researchers hope that building on her techniques may be a viable way to answer questions about sphere packing in higher dimensions. “This is the beginning of understanding sphere packings rather than the end,” says Henry Cohn, a mathematician at Microsoft Research and one of Viazovska’s collaborators for the 24-dimensional case.

Although it is virtually impossible to visualize eight-dimensional space, mathematicians are comfortable working with spaces of eight, 24 or thousands of dimensions by analogy to lower-dimensional spaces. In three dimensions points are labeled using three coordinates—length, width and height, or x, y, z—so in eight dimensions points are labeled using eight coordinates.

In three dimensions a sphere is the set of points in three-dimensional space that are all equidistant from one center point. In eight dimensions it is the set of points in eight-dimensional space that are all equidistant from one center point. In any dimension the sphere-packing problem is the question of how equal-size spheres can be arranged with as little empty space between them as possible.

Whereas it would seem logical for mathematicians to solve successively higher dimensions in turn—after solving three, researchers could build on their work to solve four and then five—it is no accident that Viazovska leapfrogged over dimensions four through seven and solved sphere packing in eight dimensions or that 24 was the one to follow that. “Part of what I love about the sphere-packing problem is that every dimension has its own idiosyncrasies,” says Cohn, who has worked at sphere packing for many years. “Some dimensions just behave much better than others.”

In two dimensions sphere—or in this case circle—packing is easy because circles of the same size fit together so snugly. Each circle can be surrounded by exactly six other circles, and there is no wiggle room. In three dimensions there is no such super-snug packing. In fact, it is not until eight dimensions that there is another configuration, called the E8 lattice packing, where everything locks into place. In 24 dimensions a packing pattern called the Leech lattice has a similar property. Such patterns are what made these dimensions so amenable to attack.

A 2003 paper by Cohn and Harvard University mathematician Noam Elkies described a new technique for finding bounds for packing densities in many different dimensions. Their approach does not directly consider packings, but rather auxiliary functions— formulas with special properties. They believed their method could be extended to a complete solution of the sphere-packing problem in eight and 24 dimensions if they could find the right functions, but the functions eluded researchers for more than a dozen years. Viazovska says she almost gave up hope but then found a function that at first looked unrelated but aligned with Cohn’s and Elkies’ work perfectly. “For me, this was the moment when everything changed and I understood the problem really could be solved,” Viazovska says.

“Many other people spent a long time looking for the function that’s needed to make that approach work, and nobody had any solid clues about

how to find it," says Hales, who knows firsthand. In addition to solving the sphere-packing problem in three dimensions, he has worked on it in other dimensions as well and spent time looking for the function himself. "I think we were all quite shocked when Maryna Viazovska made the announcement of this discovery."

Cohn says that with a problem like sphere packing the solution can provoke one of two responses: embarrassment, because the answer seems so obvious in retrospect; or awe, because the work was truly novel. Viazovska's solution is a case of the latter. "Her definitions at first glance looked kind of ad hoc. Why on Earth would you do that? But it's justified by the ingenious transformations she had," he says. "It's nice to be able to look at it and feel admiration rather than regret."

The question of how to pack spheres most tightly into higher-dimensional spaces may seem like the kind of problem only a mathematician could ever love. It turns out, however, to be far from impractical. Higher-dimensional sphere packings form the basis for error-correcting codes that help us transmit data over cellular networks, fiber-optic cables and other places where information can be lost or altered in transit. These applications treat pieces of data as points in a higher-dimensional space.

Although Viazovska's methods are not likely to solve the sphere-packing problem in other dimensions, at least not without another big breakthrough, they may help researchers improve their estimates of how tightly spheres can be packed in higher-dimensional spaces. Such advances are not as glamorous as solving the problem completely but could represent a significant improvement in higher-dimensional spaces where the data transmission stakes are high.

--Originally published: Scientific American Online June 30, 2016.

SECTION 2

The Body Mathematic

Outcalculating the Competition

by Heather Wax

In March 2008 the press went crazy for Martin A. Nowak's study on the value of punishment. A Harvard University mathematician and biologist, Nowak had signed up some 100 students to play a computer game in which they used dimes to punish and reward one another. The popular belief was that costly punishment would promote cooperation between two equals, but Nowak and his colleagues proved the theory wrong. Instead they found that punishment often triggers a spiral of retaliation, making it detrimental and destructive rather than beneficial. Far from gaining, people who punish tend to escalate conflict, worsen their fortunes and eventually lose out. "Nice guys finish first," headlines cheered.

It wasn't the first time Nowak's computer simulations and mathematics forced a rethinking of a complex phenomenon. In 2002 he worked out equations that can predict the way cancer evolves and spreads, such as when mutations emerge in a metastasis and chromosomes become unstable. And in the early 1990s his model of disease progression demonstrated that HIV develops into AIDS only when the virus replicates fast enough so that the diversity of strains reaches a critical level, one that overwhelms the immune system. Immunologists later found out he had the mechanism right. Now Nowak is out to do it again, this time by modeling the origin of life. Specifically, he is trying to capture "the transition from no life to life," he says.

Trained as a biochemist, the 43-year-old Nowak believes that mathematics is the "true language of science" and the key to unlocking the secrets of the past. He began exploring the mathematics of evolution as a graduate student at the University of Vienna, working with fellow Austrian Karl Sigmund, a leader in evolutionary game theory. Evolutionary dynamics, as Nowak named the fi eld, involves creating formulas that

describe the building blocks of the evolutionary process, such as selection, mutation, random genetic drift and population structure. These formulas track, for example, what happens when individuals with different characteristics reproduce at different rates and how a mutant can produce a lineage that takes over a population.

At the home of the Program for Evolutionary Dynamics at Harvard, the blackboard is chalked with equations. Nowak has been busy working on how to whittle down the emergence of life into the simplest possible chemical system that he can describe mathematically. He uses zeroes and ones to represent the very first chemical building blocks of life (most likely compounds based on adenine, thymine, guanine, cytosine or uracil). Nowak refers to them as monomers, which, in his system, randomly and spontaneously assemble into binary strings of information.

Nowak is now studying the chemical kinetics of this system, which means describing how strings with different sequences will grow. The fundamental principles of this idealized scheme, he says, will hold true for any laboratory-based chemical system in which monomers self-assemble, “in the same way as Newton’s equations describe how any planet goes around the sun, and it doesn’t matter what that planet is made of,” Nowak explains. “Math helps us to see what the most crucial and interesting experiment is. It describes a chemical system that can be built, and once it’s built, you can watch the origin of evolution.”

Could it really be that simple? Right now the system exists only on paper and in the computer. Although it is easy to model mathematically, making the system in the lab is tricky because it starts without any enzymes or templates to help the monomers assemble. “It’s hard to imagine an easy way to make nucleic acids,” says David W. Deamer, a biomolecular engineer at the University of California, Santa Cruz. “There had to be a starting material, but we’re very much into a murky area, and we don’t have good ideas about how to re-create it in the laboratory or how to get it to work using just chemistry and physics without the help of enzymes.”

In the 1980s biochemist Leslie E. Orgel and his group at the Salk Institute for Biological Studies in San Diego showed that a strand of RNA can act as a template for making another strand of complementary RNA—a phenomenon called nonenzymatic template-directed polymerization.

Figuring out how nucleotides might self-assemble without templates, however, has proved harder. “I want a process that can comprise polymers,” Nowak says.

Irene Chen, a cellular origins researcher at Harvard, says one way that monomers of RNA or DNA might form polymers in the absence of enzymes is by adding a compound called imidazole to one end of the monomers, making them more reactive and their polymerization quicker and easier. Lipids or clay might also be essential—other researchers have shown that they can help speed up the reaction. At Rensselaer Polytechnic Institute, for instance, chemist James P. Ferris induced adenine nucleotides to assemble into short polymers of RNA—strands 40 to 50 nucleotides long—on a kind of mineral clay that may have been common in the prebiotic world.

Using his mathematical model, Nowak looks at chemical reactions that lead to these kinds of strands and assigns rate constants to the reactions. That is, he imagines that strings with different binary information grow at different rates, with some taking in monomers faster than others. Then he calculates their distributions. Small differences in growth rates, he has noticed, result in small differences in abundance; sequences that grow slower are less common in the population, getting outcompeted by faster ones. “This I find great,” Nowak exclaims, “because now you have selection prior to replication in a completely natural way.”

Some strands mutate, and sometimes one sequence accelerates the reaction rates of other sequences, demonstrating the kind of cooperation that Nowak has long argued is a fundamental principle of evolution. Taken together, he says, the result is a lifelike chemical system ripe with evolutionary dynamics. He calls this system “prelife” because “it has the qualities of life—genetic diversity, selection and mutation—but not replication.”

Typically mutation and selection are seen as consequences of replication. If suddenly, for example, only large, hard seeds were available to the finches of the Galápagos Islands, those with bigger, stronger beaks would be more likely to survive and, generation after generation, would become more common in the population. Selection for a trait, be it beak size or something else, depends on passing down the genes for that trait to

offspring. But Nowak says his model shows there can be selection prior to replication—which means that maybe there is selection *for* replication. If this kind of selection is possible, he notes, maybe it can help explain the origin of life.

All that is necessary is for a few strings to suddenly develop the ability to make copies of themselves—the way some researchers believe certain strands of RNA first became dominant on the primitive earth. Enough free monomers would have to be around to make replication advantageous, Nowak points out, and the replicating strings must be able to use up the monomers faster than the nonreplicating strings. According to his calculations, only when the rate of replication went beyond a certain threshold would the equilibrium of the system change, allowing life to emerge. “Life destroys prelife,” he states. “All of this happened at some stage.”

Nowak hopes that his model will guide experiments. When it comes to understanding the beginning of evolution, building the chemical system he describes mathematically—a system in which only two types of monomers self-assemble and then self-replicate—“is the simplest thing you can do,” he says. “Mathematics is the proper language of evolution. I don’t know what the ‘ultimate understanding’ of biology will look like, but one thing is clear: it’s all about getting the equations right.”

--Originally published: Scientific American 299(4); 96-99 (October 2008).

The Generative Geometry of Seashells

by Derek E Moulton, Alain Goriely and Régis Chirat

Mollusks are fabulous architects. They build houses that protect their soft bodies from predators and the elements—shells of uncommon strength, durability and beauty. Many of these shells have spectacularly complex shapes—logarithmic spirals bedecked with fractal spines or other ornaments, all executed with near-perfect mathematical regularity. Yet mollusks, of course, know nothing of math. How, researchers have wondered, do these humble creatures produce such intricate patterns so precisely?

For more than 100 years scientists have recognized that cells, tissues and organs must respond to the same physical forces that govern other kinds of matter. But for most of the 20th century biologists focused on understanding how the genetic code directs the formation of biological patterns and on figuring out how those patterns function. In recent decades, however, investigators have begun to apply physics-based mathematical modeling to questions about biological form. Our own work along these lines over the past few years has yielded intriguing insights into how shells acquire their ornate structures.

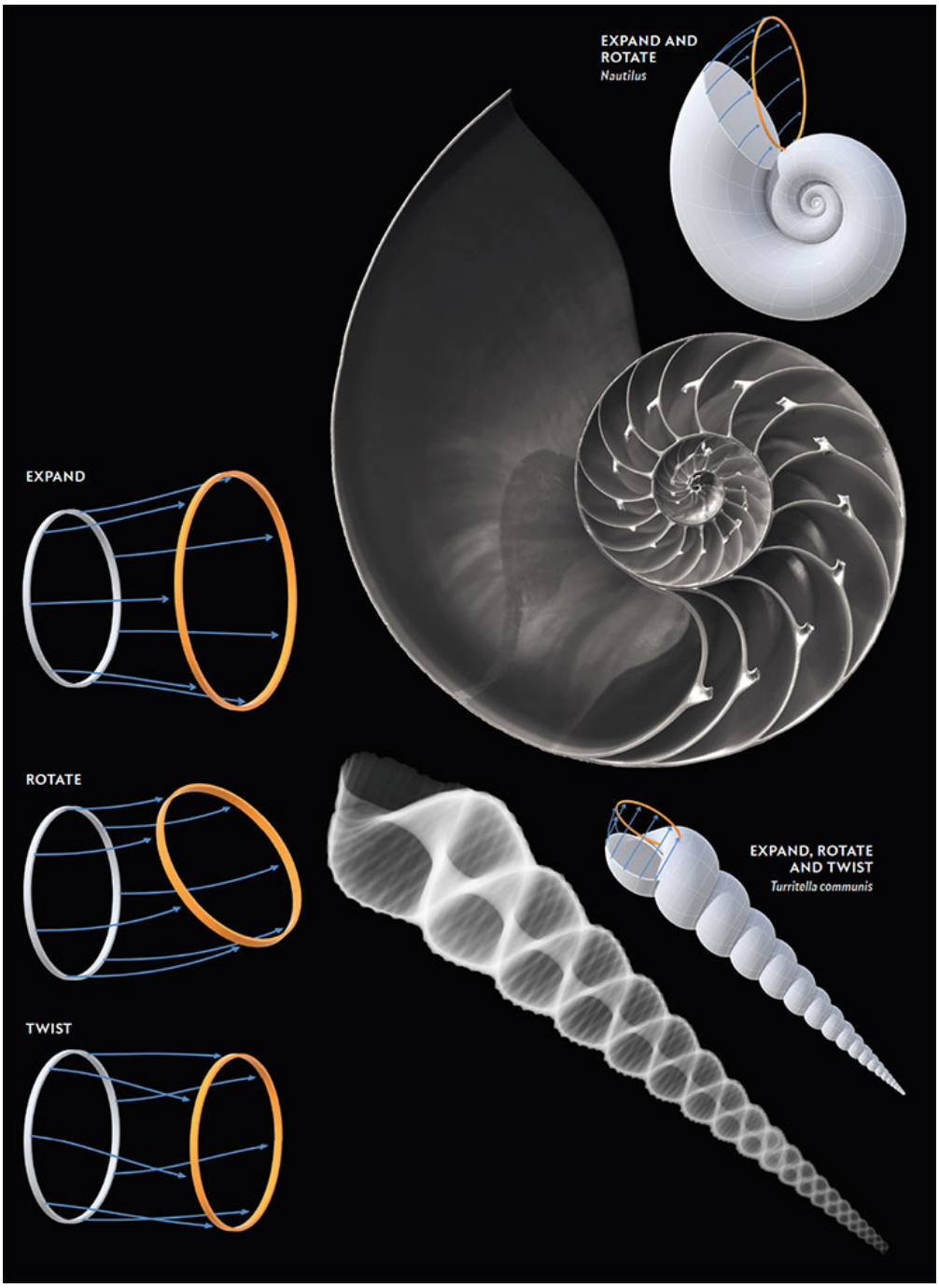
Using the tools of differential geometry, a mathematical discipline that studies curves and surfaces, we have determined that the elaborate shapes of shells arise from a few simple rules that the mollusks follow when constructing their homes. These rules interact with mechanical forces produced during shell growth to generate myriad pattern variations. Our findings help to explain how Byzantine features such as spines have evolved independently in so many lineages of gastropods, which make up the largest mollusk group. These creatures need not undergo the same genetic changes to acquire similar ornaments, because the laws of physics do most of the work.

Rules of Construction

The business of building the shell falls to the mollusk's mantle. This thin, soft organ secretes layer on layer of a substance rich in calcium carbonate at the opening, or aperture, of the shell. It only needs to follow three basic rules to form the characteristic spiral seen in the shells of snails and their relatives, the gastropods. The first rule is expand: by uniformly depositing more material than it did on the previous pass, the mollusk creates a slightly larger opening at each iteration. This process generates a cone from an initial circle. The second rule is rotate: by depositing slightly more material on one side of the aperture, the mollusk achieves a full rotation of that aperture, building a doughnut shape, or torus, from an initial circle. The third rule is twist: the mollusk rotates the points of deposition. Follow just the expand and rotate operations, and you get a planospiral shell like that of the chambered nautilus. Add the twist step, and the result is what mathematicians describe as a nonplanar, helicospiral shell.

Spirals

Mollusks follow just a few simple rules to create spiral shells. The first is expand: as the mollusk secretes successive layers of shell-building material at the shell opening, or aperture, it uniformly deposits more material each time to create an ever larger opening. The second rule is rotate: by depositing slightly more material on one side of the aperture, the mollusk builds a doughnut shape, or torus, from an initial circle. The third rule is twist: the mollusk rotates the points of deposition. Different combinations of these rules yield different spiral shapes.



Illustrations: Bryan Christie Design; Images: Ben Brain Getty Images (Nautilus); Nick Veasey Getty Images (Turritella communis)

For some shell builders, that is the end of the story, as sleek and elegant a home as one could want. For others, some embellishment is in order. To understand how ornaments such as spines form, we must examine the forces produced during shell growth. The shell secretion process revolves around an interesting mechanical system. The mantle is attached to the shell via the so-called generative zone, a region of secreted but not yet calcified

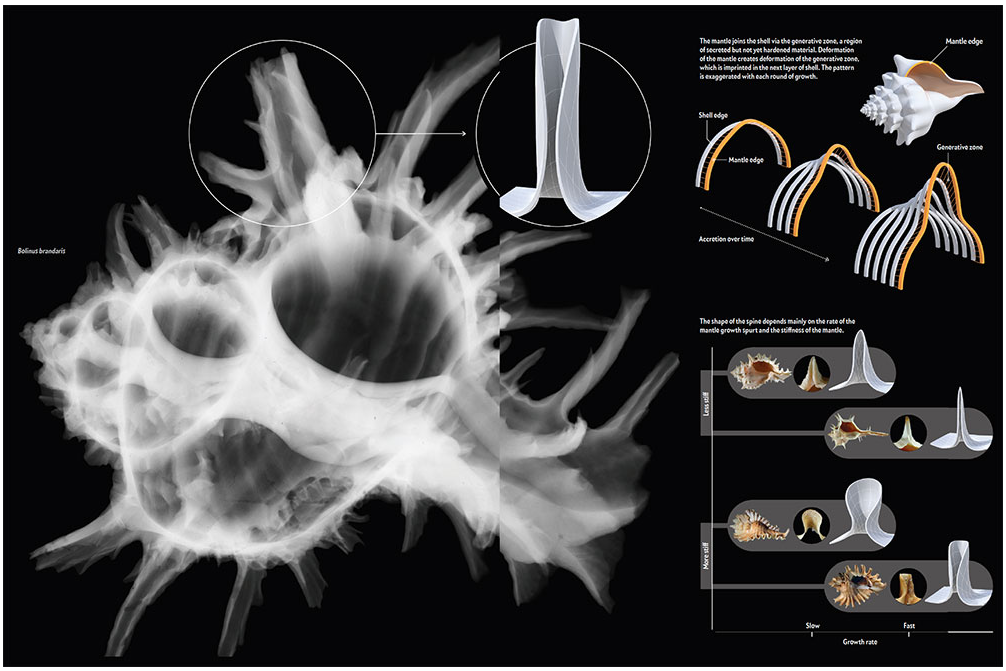
material. It is in this interaction between mantle and shell that the potential for pattern formation exists. Any mismatch between mantle and aperture will physically stress the mantle tissue. If the mantle is too small for the opening, it will have to stretch to attach to it. If the mantle is too large, it will have to compress to fit. And if the generative zone becomes deformed be cause of these stresses, the new material the mantle secretes at that stage will assume the deformed shape and permanently solidify in the shell, further influencing the mantle at the next growth step. Essentially, if the shell does not grow at the exact same rate as the growing mollusk, deformations will arise, generating features we recognize as ornaments.

Spines constitute the most prominent ornamentation, typically protruding at a right angle to the shell aperture and often extending centimeters beyond the shell surface. These projections form at regular periods in which the mantle undergoes a growth spurt. During a growth spurt, the mantle develops so quickly that it has an excess of length and cannot align with the aperture. This mismatch causes the mantle to buckle slightly. The material it secretes assumes the buckled shape. By the next increment, the mantle has grown further and has again exceeded the aperture, which has the effect of amplifying the buckled pattern. We reasoned that this repeated process of growth and mechanical interaction gives rise to a row of spines, the precise pattern of which is determined primarily by the rate of the growth spurt and the stiffness of the mantle.

To test this idea, we developed a mathematical model of a mantle growing on a foundation that evolves at each iteration. As we experimented with typical growth and material properties in the model, a wide variety of spine patterns emerged, similar to the forms that are ob served in real shells, confirming our hypothesis.

Spines

An organ called the mantle is responsible for secreting the substance that becomes the shell. Spines form at regular periods during mantle growth spurts, when the organ expands so quickly that it cannot align with the aperture. This mismatch causes the mantle to buckle slightly. The shell-building material it releases then assumes the buckled shape. Each round of mantle growth and subsequent mechanical conflict with the aperture amplifies the buckled pattern.



Illustrations: Bryan Christie Design; Images: Ben Brain Getty Images (Nautilus); Nick Veasey Getty Images (Turritella communis)

This Old House

Spines are not the only flourish that mollusks may add to their shells. Another type of pattern is found on the shells of ammonites, a group of extinct mollusks related to today's cephalopods (nautiluses, octopuses and their cousins). Ammonites ruled the seas for 335 million years before disappearing around 65 million years ago. The abundance of their fossilized remains, along with their great diversity of forms and apparently high rate of evolution, has made them one of the most studied groups of fossil invertebrates.

The most striking characteristic of the ammonite shell beyond its planar-logarithmic spiral form is the regular ribbing that runs parallel to the shell edge. This ornamentation probably stems from the same mechanical conflict that produces spines, yet it is a completely different pattern. The forces are the same, but the magnitude and the geometry on which they operate are not.

The aperture of the ammonite is basically circular. If the mantle radius is larger than the current aperture radius, the mantle will be compressed but not enough to generate the degree of elastic instability needed to produce spines. Rather the compressed mantle pushes outward, and the shell radius

at the next increment is larger. But this outward motion is opposed by the calcifying generative zone, which acts as a torque spring trying to maintain the current shell orientation.

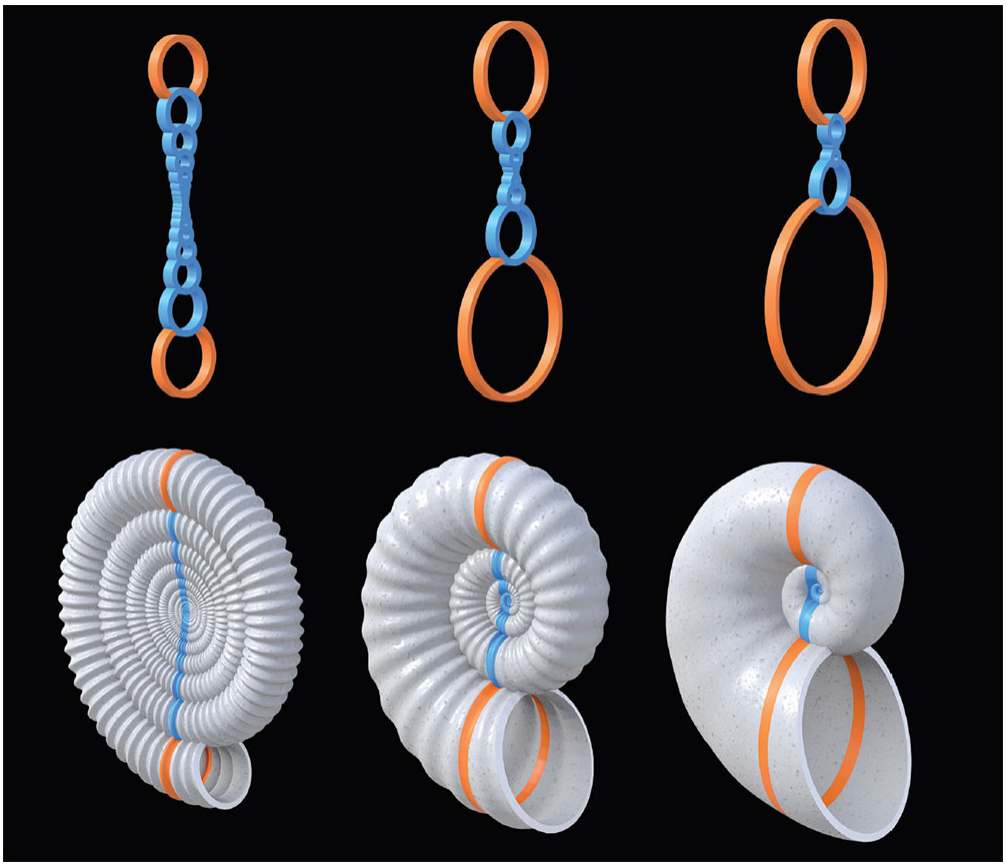
We surmised that the effect of these two opposing forces is an oscillatory system: the shell radius increases, reducing compression, but overshoots to a state of tension; the “stretched” mantle then pulls in ward to decrease its tensile force, again overshooting to a state of compression. A mathematical description of this “morphomechanical oscillator” confirmed our hypothesis, producing regular ribs with a wavelength and amplitude that increased during the growth and development of the mollusk. These mathematical predictions closely resemble the known forms of ammonites.

Mathematical modeling also predicts that the greater the expansion rate of the growing mollusk—the rate at which the diameter of the shell opening in creases—the less pronounced its ribs are. These findings help to explain the observation that increased aperture curvature correlates with increased ribbing pattern, an evolutionary trend that has been noted by paleontologists for more than a century.

This relation between expansion rate and ribbing also provides a simple mechanical and geometric explanation for a long-standing puzzle of mollusk evolution: the shells of the chambered nautilus and its relatives—a group known as the nautilids—have remained essentially smooth since at least 200 million years ago, leading some observers to suggest that the group has apparently not evolved in that time. Indeed, today’s few surviving nautilid species are often described as “living fossils.” Our biophysical growth model, however, shows that the smoothness of nautilid shells is merely a mechanical consequence of rapid aperture expansion. The nautilids’ lineage may have evolved more than their shell morphology suggests, but lacking the distinctive ornamental patterns that paleontologists use to distinguish species, their actual evolution remains largely hidden.

Ribs

Shells of ammonites, a group of extinct mollusks, exhibit regular ribs that form perpendicular to the shell edge. Mathematical modeling indicates that this pattern of ornamentation is the product of the opposing forces of the mantle and the generative zone, which form an oscillatory system of tension and compression. Slow expansion of the mollusk aperture leads to dense ribbing (left), whereas rapid expansion leads to smooth shells (right).



Illustrations: Bryan Christie Design; Images: Ben Brain Getty Images (Nautilus); Nick Veasey Getty Images (Turritella communis)

We still have much to learn about how mollusks make their marvelous abodes. A short stroll through any good shell collection reveals a number of patterns scientists have yet to explain. For example, roughly 90 percent of gastropods are “right-handed,” building their shells such that they coil in a clockwise direction. Only 10 percent wind to the left. Scientists have only just begun to probe the mechanisms that lead to this prevalence of right-handedness. The origins of some exquisite ornamentations are likewise un known, such as the fractallike spine pattern found in a number of species in the muricid family of mollusks. Also, although we know that environmental factors influence shell growth rate, the impact of these variables on shell form is less clear.

With these and other mysteries still surrounding seashells—which are model organisms for exploring broader questions about pattern formation in nature—we have our work cut out for us. But an understanding of the physical forces that govern their development only heightens their allure.

--Originally published: Scientific American 318(4); 68-75 (April 2018).

Modeling the Flu

by Adam J. Kucharski

When it comes to infectious diseases, children get a tough deal. Not only do they spend all day in a school-shaped mixing pot of viruses and bacteria, they do not yet have the repertoire of immune defenses their parents have spent a lifetime building—which means that for most infections, from chickenpox to measles, it pays to be an adult.

Influenza is a different story, however. Studies of the 2009 flu pandemic have shown that immunity against regular seasonal flu viruses tends to peak in young children, drop in middle-aged people and then rise again in the elderly. Adults might have had more exposure to the disease in the course of their lives, but—aside from the eldest group—they somehow end up with a much weaker immune response.

This curious observation naturally leads biologists to wonder about the causes. Understanding influenza infection is far from straightforward, but we are starting to find some clues in mathematical models that simulate the immune system. These models allow us to explore how past exposure to flu viruses might influence later immunological responses to new infections and how the level of protection could change with age. By bringing together these mathematical techniques with observed data, we are beginning to unravel the processes that shape immunity against influenza. In the process, the work provides new support for a quirky hypothesis—first proposed more than half a century ago and known as original antigenic sin—about why the body's response to this illness is biased toward viruses seen in childhood. Taking these insights into account is already helping us to understand why some populations suffered so unexpectedly badly in past outbreaks and might eventually help us anticipate how different groups of people will react to future outbreaks, too.

A Model Epidemic

To date, most mathematical models of immunity have not looked at the body's reaction to the influenza virus, because the pathogen is so variable. Historically, models have instead focused on the response to viruses such as measles, which change so little over time that they trigger lifelong immunity. Once individuals recover from measles or are vaccinated against it, the immune system promptly recognizes the proteins on the surface of the virus, generates antibody molecules targeted against those proteins and homes in on them to neutralize any subsequent interlopers. (Scientists call these surface proteins "antigens," an abbreviation of *anti*body *gen*erator.)

If people have a certain probability of getting infected with measles every year, one might expect immunity (measured by testing the potency of an individual's antibodies in the blood) to gradually increase with advancing years—as has been observed in several laboratory studies across differing age groups. One way to test such an explanation is to use a mathematical model, which can show what patterns one might expect to see if a theory were true. Models are powerful tools because they allow us to examine the effects of biological processes that could be difficult or even unethical to reproduce in real experiments. For example, we can see how infection might influence immunity in a population without having to deliberately infect people.

In the simplest epidemic model, a population is divided into three compartments: people who are susceptible to an infection, those who have become sick and those who have recovered from—and are therefore immune to—the disease. During the 1980s epidemiologist Roy M. Anderson, zoologist Robert M. May and their colleagues used such models to examine the age distribution of immunity to a disease such as measles. Although a three-compartment model reproduced the general pattern, they found that real-world immunity increased at a faster rate in younger age groups than the model led them to expect. Perhaps the discrepancy occurred because children had more contacts with others and thus more exposures than did those in older age groups? By updating their model to include this variation, the researchers could test the prediction. Indeed, when they altered their calculations so that children were given a higher risk of

infection, it was possible to re-create the observed changes in immunity with age.

Unfortunately, immunity against influenza is not so straightforward. Flu viruses have a high rate of mutation, which means their antigens can change appearance from year to year. As a result, the body can struggle to recognize a new strain. This variability is why flu vaccines need to be updated every few years; unlike the measles virus, which looks the same every year, antigens from the flu virus change over time.

When I first became aware of the unusual age distribution of flu immunity in the 2009 data, I wondered whether the high rate of mutation for flu virus—along with intense social contact between children—could explain the rise-dip-rise pattern across age groups. Be cause people are exposed to lots of infections when they are young, they are likely to develop good, long-term immunity against the bulk of viruses that circulated during their childhood. In the case of flu, children do develop antibodies against the antigens of specific influenza viruses they meet, just as they do for measles.

After leaving high school or college, however, folks meet fewer people on average and so will generally catch the flu less frequently. This change in exposure means adults rely on the antibodies they built up as children to protect them against any new assaults. Yet because flu viruses change over time, their “old” antibodies would be less effective with advancing years at recognizing newer strains. Hence, one might expect levels of natural protection to drop in middle-aged adults—who, as a group, do not receive routine flu immunizations. And the subsequent rise in immunity seen in elderly individuals might occur because they often receive flu shots, which keep their antibodies up-to-date.

That was the theory, at least. The problem was how to test it. Because flu is so variable, it is much harder to build a mathematical model for it than for measles. Even if a person is immune to one strain, he or she might be only partially immune to another and completely susceptible to a third. To study immunity, we therefore need to keep precise track of the combination of influenza strains to which people have been exposed and in what order the exposures occurred.

This is where it gets tricky because of the vast number of combinations of strains that people could have seen. If 20 different strains have circulated in the past, for example, there would be 2^{20} (or more than one million) possible histories of infection for any particular individual. For 30 strains, there would be more than one billion combinations for each individual.

Along with Julia R. Gog, then my Ph.D. supervisor at the University of Cambridge, I set out to find a way around this mountain of complexity. We realized that if individuals had a certain probability of becoming exposed to flu every year, the probabilities of coming into contact with any two strains should be independent of each other. (In other words, exposure to strain A should not affect the chances of being exposed to strain B.) Thus, for fundamental mathematical reasons, we could reconstruct the probability that a random individual had been exposed to a certain combination of infections simply by multiplying the probabilities of exposure to each individual strain in the combination. This meant that instead of dealing with one million probabilities for 20 different strains, we would have to deal with only 20.

When we ran the equations for the model, however, the results were not what we expected. The model stubbornly suggested that if a person had previously been exposed to even a single strain, he or she was *more* likely to have seen another one. It was as if our model was saying that being hit by lightning made you more likely to have been exposed to flu—an obviously absurd conclusion.

The reason for this seemingly nonsensical result turned out to be simple: we had not accounted for a person’s age. Assuming infections occur at a fairly consistent rate, the longer a person is alive, the more likely it is that the individual will contract at least one infection. So if you pick a random individual—say, a female—and learn she was previously exposed to flu (or was struck by lightning), you immediately know she is more likely to be older than younger. And because she is older, you know that she is more likely to have experienced some other misfortune—such as exposure to a second flu strain.

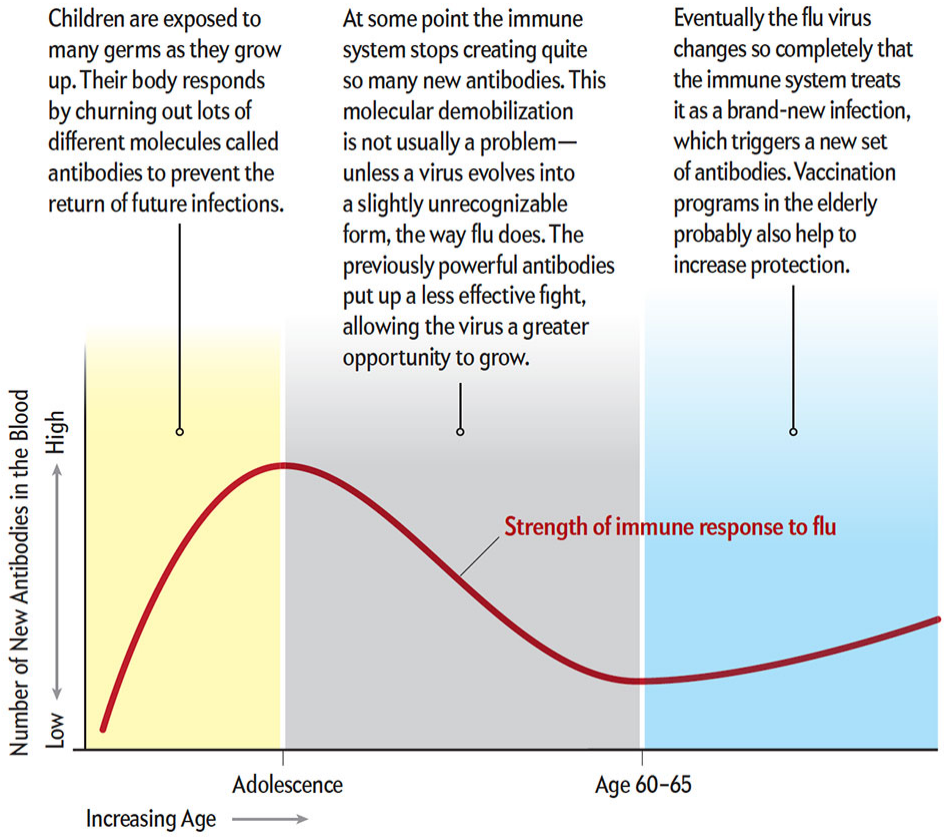
As long as we dealt with each age group separately, however, the number of infections went back to being independent variables. Thus, for 20 strains, we no longer had one million things to keep track of: we were back to

having only 20. With a viable model in place, we started to build simulations of how the body's immunity to influenza changed over time. The aim was to generate artificial data that we could test against real-life patterns. As well as having the virus mutate over the years, we assumed that each age group's risk of infection depended on the number of social contacts reported in population surveys within and between different age groups.

Alas, even with these changes, our model—which assumed that the middle-age dip in immunity arose from fewer exposures—could not reproduce the midlife drop seen in the real world. The model was not completely incorrect: it showed that children developed a stronger immunity than adults. But whereas the actual drop off in antibody levels appears to start between five and 10 years of age, in our model the decline occurred between 15 and 20 years of age—after individuals would have left school (where there are lots of people and germs).

First Impressions Last a Long Time

Most of the time when the human body conquers a virus, the immune system can provide lifelong protection against future infections with the same pathogen. Adults should therefore have stronger defenses than children and become sick less often. But that is not what happens with flu. Immunity grows throughout childhood, as expected, but then people become more vulnerable in middle age (curve). A possible explanation appears below: the immune system develops a kind of blind spot with flu, mistakenly expecting later infections of the highly changeable virus to resemble earlier ones. Because the body reserves its strongest responses for what are in fact outdated threats, it can fail to combat subsequent infections effectively.



Source: "The Role of Social Contacts and Original Antigenic Sin in Shaping the Age Pattern of Immunity to Seasonal Influenza," by Adam J. Kucharski and Julia R. Gog, in PLOS Computational Biology, Vol. 8, No. 10, Article No. E1002741; October 25, 2012.

Original Sin

While puzzling over the flu age pattern, I had talked to many people about the wider problem of modeling immunity. In particular, I spoke with Andrea Graham, an evolutionary biologist at Princeton University, who introduced me to the concept of original antigenic sin. Now that we had a model that could handle a large number of strains, I wondered if taking this hypothesis into account would help our model produce more realistic results. Because the idea was controversial, I also wondered if incorporating it might help indicate whether it was plausible or not.

Like the biblical concept, original antigenic sin is the story of the first encounter between a naive entity (the immune system) and a dangerous threat (a pathogen). In the immunological version, the body is so marked by its first successful counterattack against an influenza virus that each subsequent infection will trigger these original antibodies again. The body makes these antibodies even when it encounters a slightly different set of

antigens on a pathogen, which would require a different set of antibodies for the host to combat the infection efficiently. At the same time, the body fails to make a good supply of antibodies against the pathogen with the altered set of antigens, instead relying on the immune response to viruses it has already seen.

Virologist Thomas Francis, Jr., first came across the problem in 1947. Despite a large vaccination program in the previous year, students at the University of Michigan had fallen ill with a new, albeit related, influenza strain. When Francis compared immunity against the vaccine strain with immunity against the new virus, he found that the students possessed antibodies that could target the vaccine strain effectively but not the virus with which they had been infected a year later.

Eventually Francis developed an explanation for his curious observation. He suggested that instead of developing antibodies to every new virus that it encountered, the immune system might reproduce the same reaction to similar viruses it had already seen. In other words, past strains and the order in which people get them could be very important in determining how well a person could fight off subsequent outbreaks of the ever variable flu virus. Francis called the phenomenon “original antigenic sin”—perhaps, as epidemiologist David Morens and his colleagues later suggested, “in religious reverence for the beauty of science or impish delight fueled by the martini breaks of which he was so fond.”

During the 1960s and 1970s researchers found further evidence of original antigenic sin in humans and other animals. Since then, however, other studies have questioned its existence. In 2008 researchers at Emory University and their colleagues examined antibody levels in volunteers who had received flu shots and found that their immune system was effective at targeting the virus strain in the vaccine. The researchers’ concluded that original antigenic sin “does not seem to be a common occurrence in normal, healthy adults receiving influenza vaccination.” The following year, however, another Emory-based group, led by immunologist Joshy Jacob, found that full-scale infection in mice with a live flu virus—rather than an inactivated virus, as is typically present in a vaccine—could hamper subsequent immune responses to other strains, suggesting anew that

original antigenic sin may play a more important role during natural infections with flu.

Jacob and his group proposed a biological explanation for original antigenic sin, hypothesizing that it could stem at root from how we generate so-called memory B cells. These cells form part of the immune response: during an infection they are programmed to recognize a specific threat and produce antibodies that finish it off. Some B cells persist in the body after a siege, ready to spew more antibodies should the same threat reappear. According to Jacob and his colleagues, infection with live influenza viruses could trigger existing memory cells to action rather than causing new B cells to be programmed. Suppose you were infected with flu last year and then catch a slightly different virus this year. Because memory B cells have already seen last year's similar virus, they can get rid of it before the body has time to develop new B cells that are specific to—and hence better at remembering—this year's strain. It is like the old military adage about generals always fighting the last war (especially if they won it). It seems the immune system depends more on shoring up past defenses rather than generating new ones, especially if the old strategy works reasonably well and more quickly.

During the final stages of my Ph.D., we adapted our new model to simulate original antigenic sin. This time the distinctive decline in immunity showed up in our simulation right when it does in real life—after about age seven, when people are old enough to have seen at least one flu infection (instead of between ages 15 and 20). From that point onward, our model suggested, previous infections compromised the creation of effective antibodies. (Because younger individuals in the countries we studied are not typically vaccinated, this effect is likely to come from natural infection with flu.) It is still not completely clear what causes the increase in immunity in the eldest group. It could be partly the result of increased vaccination in that age range or partly the fact that individuals have been alive so long that the antigens of any new flu strains to which they are exposed are so different that they can no longer be mistaken by the immune system for the viruses from childhood. At any rate, our findings suggested that original antigenic sin, rather than the number of social contacts (and thus chances of exposure), was responsible for the curious age distribution of immunity in younger people.

Blind Spots

Having become convinced that original antigenic sin can shape the immune profile of an entire population, we wanted to investigate whether misguided immune responses could also affect the size of an outbreak. In simulations, we found that every now and then, the model generated large epidemics even if the new virus was not particularly different from the previous year's strain. It seemed that original antigenic sin was leaving gaps in the immunity of certain age groups: although individuals had been exposed to strains that might have protected them, their immune systems had generated the "wrong" antibodies in response to the new infection.

The best historical evidence supporting this idea came from 1951, when influenza rippled across the English city of Liverpool in a wave that was quicker and deadlier there than the infamous "Spanish flu" pandemic of 1918. Even the two subsequent flu pandemics, in 1957 and 1968, would pale in comparison. Yet it is not clear what caused the outbreak to be so bad.

The most logical explanation was that the 1951 strain must have been very different from the strain circulating in 1950 and that, therefore, most people would not have had an effective immune response when the virus hit them. But there is not much evidence that the 1951 strain was significantly different from the one that circulated the year before. What is more, the size of the epidemic in the U.K. and elsewhere varied depending on location. Some places, such as England (particularly Liverpool) and Wales, were hit hard, whereas others, such as the U.S., saw little change in mortality from previous years. More recently, the U.K. experienced severe flu epidemics in 1990 and 2000, again without much evidence that the virus was particularly different in those years.

Yet our mathematical model could re-create conditions similar to the flu outbreaks of 1951, 1990 and 2000. When original antigenic sin was assumed to occur, the order in which different flu strains caused illness in a particular age group could shape how well its members fought off future flu infections. In other words, when it comes to flu, each geographical location may have its own unique immune profile, subtly different from its neighbors, with its own unique "blind spots" in immunity. Severe outbreaks such as the one in Liverpool may therefore have been the caused by such

blind spots, which other regions simply did not have, because they experienced a different original antigenic sin.

Refining Original Antigenic Sin

Research into influenza immunity has often focused on specific issues, namely the effectiveness of a particular vaccine or the size of an epidemic in a certain year. But these problems are actually just part of a much bigger question: How do we develop and maintain immunity to flu and other viruses that change their antigenic makeup over time—and can we use that information to understand how flu spreads and evolves?

Projects such as the FluScape study in southern China are now starting to tackle the problem. A preliminary analysis published in 2012 by Justin Lessler of the Johns Hopkins Bloomberg School of Public Health and his colleagues suggested that the concept of original antigenic sin might need to be refined. Rather than the immune response being dictated only by the first strain an individual encountered, the researchers found evidence that immunity follows a hierarchy. They suggested that the first strain someone was infected with gained the most “senior” position in the immune response, with the next strain generating a somewhat weaker response, followed by an even weaker response for the third strain. (Such a seniority hierarchy would apply only to highly variable viruses, such as flu viruses.)

Because the FluScape study looked at blood samples taken in the present day, Lessler and his colleagues could not examine how antibody levels changed over time. In August 2013, however, researchers at the Icahn School of Medicine at Mount Sinai looked at a series of blood samples taken from 40 people over a 20-year period. Their results support the idea of antigenic seniority: each new flu infection boosted antibody levels against previously seen strains. Individuals therefore had stronger immune responses against viruses they came across earlier in life than against those encountered later.

Over the past couple of years I have been collaborating with the FluScape team to investigate patterns in the new data coming out of China. One benefit of such work might be to help determine who is susceptible to particular strains and how this vulnerability could influence the evolution of the disease. With new models and better data, we are gradually starting to

find ways to tease out how individuals and populations build immunity to influenza. If the past is anything to go by, we are sure to encounter more surprises along the way.

--Originally published: Scientific American 311(6); 80-85 (December 2014).

Cracking the Brain’s Enigma Code

by Helen Shen

Brain-controlled prosthetic devices have the potential to dramatically improve the lives of people with limited mobility resulting from injury or disease. To drive such brain-computer interfaces, neuroscientists have developed a variety of algorithms to decode movement-related thoughts with increasing accuracy and precision. Now researchers are expanding their tool chest by borrowing from the world of cryptography to decode neural signals into movements.

During World War II, codebreakers cracked the German Enigma cipher by exploiting known language patterns in the encrypted messages. These included the typical frequencies and distributions of certain letters and words. Knowing something about what they expected to read helped British computer scientist Alan Turing and his colleagues find the key to translate gibberish into plain language.

Many human movements, such as walking or reaching, follow predictable patterns, too. Limb position, speed and several other movement features tend to play out in an orderly way. With this regularity in mind, Eva Dyer, a neuroscientist at the Georgia Institute of Technology, decided to try a cryptography-inspired strategy for neural decoding. She and her colleagues published their results in a recent study in *Nature Biomedical Engineering*.

“I’ve heard of this approach before, but this is one of the first studies that’s come out and been published,” says Nicholas Hatsopoulos, a neuroscientist at the University of Chicago, who was not involved in the work. “It’s pretty novel.”

Existing brain-computer interfaces typically use so-called ‘supervised decoders.’ These algorithms rely on detailed moment-by-moment

movement information such as limb position and speed, which is collected simultaneously with recorded neural activity. Gathering these data can be a time-consuming, laborious process. This information is then used to train the decoder to translate neural patterns into their corresponding movements. (In cryptography terms, this would be like comparing a number of already decrypted messages to their encrypted versions to reverse-engineer the key.)

By contrast, Dyer's team sought to predict movements using only the encrypted messages (the neural activity) and a general understanding of the patterns that pop up in certain movements. Her team trained three macaque monkeys to either reach their arm or bend their wrist to guide a cursor to a number of targets arranged about a central point. At the same time, the researchers used implanted electrode arrays to record the activity of about 100 neurons in each monkey's motor cortex, a key brain region that controls movement.

Over the course of many experimental trials, researchers gathered statistics about each animal's movements, such as the horizontal and vertical speed. A good decoder, Dyer says, should find corresponding patterns buried in the neural activity that map onto patterns seen in the movements. To find their decoding algorithm, the researchers performed an analysis on the neural activity to extract and pare down its core mathematical structure. Then they tested a slew of computational models to find the one that most closely aligned the neural patterns to the movement patterns.

When the researchers used their best model to decode neural activity from individual trials, they were able to predict the animals' actual movements on those trials about as well as some basic supervised decoders. "It's a very cool result," says Jonathan Kao, a computational neuroscientist at the University of California, Los Angeles, who was not involved in the study. "My prior thought would have been that having the moment-by-moment information of the precise reach, knowing the velocity at every moment in time, would have allowed you to build a better decoder than if you just had the general statistics of reaching."

Because Dyer's decoder only required general statistics about movements, which tend to be similar across animals or across people, the researchers were also able to use movement patterns from one monkey to

decipher reaches from the neural data of another monkey—something that is not feasible with traditional supervised decoders. In principle, this means that researchers could reduce the time and effort involved in collecting meticulously detailed movement data. Instead they could acquire the information once and reuse or distribute those data to train brain-computer interfaces in multiple animals or people. “It could be very useful to the scientific community and to the medical community,” Hatsopoulos says.

Dyer calls her work a proof of concept for using cryptographic strategies to decode neural activity and notes that much more work must be done before the method can be used widely. “By comparison to state-of-the-art decoders, this is not yet a competitive method,” she says. The algorithm could potentially be strengthened by feeding it signals from even more neurons or providing additional known features of movements, such as the tendency of animals to produce smooth motions. To be practical for guiding prosthetic devices, the approach would also have to be adapted to decode more complex, natural movements—a nontrivial task. “We’ve only kind of scratched the surface,” Dyer says.

--Originally published: Scientific American Mind 29(3); (March 2018).

SECTION 3

Mathematics for Understanding the Physical World

How Einstein Discovered Reality

by Walter Isaacson

The general theory of relativity began with a sudden thought. It was late 1907, two years after the “miracle year” in which Albert Einstein had produced his special theory of relativity and his theory of light quanta, but he was still an examiner in the Swiss patent office. The physics world had not yet caught up with his genius. While sitting in his office in Bern, a thought “startled” him, he recalled: “If a person falls freely, he will not feel his own weight.” He would later call it “the happiest thought in my life.”

The tale of the falling man has become an iconic one, and in some accounts it actually involves a painter who fell from the roof of an apartment building near the patent office. Like other great tales of gravitational discovery—Galileo dropping objects from the Leaning Tower of Pisa and the apple falling on Isaac Newton’s head—it was embellished in popular lore. Despite Einstein’s propensity to focus on science rather than the “merely personal,” even he was not likely to watch a real human plunging off a roof and think of gravitational theory, much less call it the happiest thought in his life.

Einstein soon refined his thought experiment so that the falling man was in an enclosed chamber, such as an elevator, in free fall. In the chamber, he would feel weightless. Any objects he dropped would float alongside him. There would be no way for him to tell—no experiment he could do to determine—if the chamber was falling at an accelerated rate or was floating in a gravity-free region of outer space.

Then Einstein imagined that the man was in the same chamber way out in space, where there was no perceptible gravity, and a constant force was pulling the chamber up at an accelerated rate. He would feel his feet pressed to the floor. If he dropped an object, it would fall to the floor at an

accelerated rate—just as if he stood on Earth. There was no way to make a distinction between the effects of gravity and the effects of being accelerated.

Einstein dubbed this “the equivalence principle.” The local effects of gravity and of acceleration are equivalent. Therefore, they must be manifestations of the same phenomenon, some cosmic field that accounts for both acceleration and gravity.

It would take another eight years for Einstein to turn his falling-man thought experiment into the most beautiful theory in the history of physics. He would go from his sedate life as a married father working at the Swiss patent office to living alone as a professor in Berlin, estranged from his family and increasingly alienated from his Prussian Academy of Sciences colleagues there by the rise of anti-Semitism. The decision last year by the California Institute of Technology and Princeton University to put an archive of Einstein’s papers online for free permits a glimpse of him juggling the cosmic and the personal throughout this period. We can relish his excitement in late 1907 as he scribbled down what he called “a novel consideration, based on the principle of relativity, on acceleration and gravitation.” Then we can sense his grumpy boredom, a week later, as he rejected an electric company’s patent application for an alternating-current machine, calling the claim “incorrectly, imprecisely and unclearly prepared.” The coming years would be full of human drama, as Einstein raced against a rival to give mathematical expression to relativity while struggling with his estranged wife over money and his right to visit his two young boys. But by 1915 his work climaxed in a completed theory that would change our understanding of the universe forever.

Bending Light

For almost four years after positing that gravity and acceleration were equivalent, Einstein did little with the idea. Instead he focused on quantum theory. But in 1911, when he had finally breached the walls of academia and become a professor at the German Charles-Ferdinand University in Prague, he turned his attention back to coming up with a theory of gravity that would help him generalize special relativity—the relation between space and time that he defined in 1905.

As Einstein developed his equivalence principle, he realized that it had some surprising ramifications. For example, his chamber thought experiment indicated that gravity would bend light. Imagine that the chamber is being accelerated upward. A light beam comes in through a pinhole on one wall. By the time it reaches the opposite wall, the light is a little closer to the floor because the chamber has shot upward. And if you could plot the beam's trajectory across the chamber, it would be curved because of the upward acceleration. The equivalence principle says that this effect should be the same whether the chamber is accelerating upward or is resting still in a gravitational field. In other words, light should bend when passing through a gravitational field.

In 1912 Einstein asked an old classmate to help him with the complicated mathematics that might describe a curved and warped four-dimensional spacetime. Until then, his success had been based on his talent for sniffing out the underlying physical principles of nature. He had left to others the task of finding the best mathematical expressions of those principles. But now Einstein realized that math could be a tool for discovering—and not merely describing—nature's laws.

Einstein's goal as he pursued his general theory of relativity was to find the mathematical equations describing two interwoven processes: how a gravitational field acts on matter, telling it how to move, and how matter generates gravitational fields in spacetime, telling spacetime how to curve.

For three more years Einstein wrestled with drafts and outlines that turned out to have flaws. Then, beginning in the summer of 1915, the math and the physics began to come together.

Personal Unraveling

By then, he had moved to Berlin to become a professor and member of the Prussian Academy. But he found himself working pretty much without support. Anti-Semitism was rising, and he formed no coterie of colleagues around him. He split with his wife, Mileva Marić, a fellow physicist who had been his sounding board in formulating special relativity in 1905, and she moved back to Zurich with their two sons, ages 10 and four. He was having an affair with his cousin Elsa, whom he would later marry, but he lived by himself in a sparsely furnished apartment in central Berlin, where

he ate intermittently, slept randomly, played his violin and waged his solitary struggle.

Throughout 1915 his personal life began to unravel. Some friends were pressing him to get a divorce and marry Elsa; others were warning that he should not be seen with her or let her come near his two boys. Marić repeatedly sent letters requesting money, and at one point Einstein replied with unbridled bitterness. “I find such a demand beyond discussion,” he responded. “I find your constant attempts to lay hold of everything that is in my possession absolutely disgraceful.” He tried hard to maintain a correspondence with his sons, but they rarely wrote back, and he accused Marić of not delivering his letters to them.

Yet amid this personal turmoil, Einstein was able to devise, by late June 1915, many elements of general relativity. He gave a weeklong series of lectures at the end of that month on his evolving ideas at the University of Göttingen in Germany, the world’s preeminent center for mathematics. Foremost among the geniuses there was David Hilbert, and Einstein was particularly eager—perhaps too eager, it would turn out—to explain all the intricacies of relativity to him.

A Rivalry

The visit to Göttingen was a triumph. A few weeks later Einstein reported to a scientist friend that he “was able to convince Hilbert of the general theory of relativity.” In a letter to another colleague, he was even more effusive: “I am quite enchanted with Hilbert!”

Hilbert was likewise enchanted with Einstein and with his theory, so much so that he soon set out to see if he could do what Einstein had so far not accomplished: produce the mathematical equations that would complete the formulation of general relativity.

Einstein began hearing Hilbert’s footsteps in early October 1915, just as he realized that his current version of the theory—which was based on an *Entwurf*, or outline, he had been refining for two years—had serious flaws. His equations did not account properly for rotating motion. In addition, he realized that his equations were not generally covariant, meaning that they did not really make all forms of accelerated and nonuniform motion relative, nor did they fully explain an anomaly that astronomers had

observed in the orbit of the planet Mercury. Mercury's perihelion—its point of closest approach to the sun—had been gradually shifting in a way not accounted for by Newtonian physics or by Einstein's then current version of his own theory.

Einstein faced two ticking clocks: he could sense that Hilbert was closing in on the correct equations, and he had agreed to give a series of four formal Thursday lectures on his theory in November to the members of the Prussian Academy. The result was an exhausting monthlong whirlwind during which Einstein wrestled with a succession of equations, corrections and updates that he rushed to complete.

Even as he arrived at the grand hall of the Prussian State Library on November 4 to deliver the first of his lectures, Einstein was still wrestling with his theory. "For the last four years," he began, "I have tried to establish a general theory of relativity." With great candor, he detailed the problems he had encountered and admitted that he still had not come up with equations that fully worked.

Einstein was in the throes of the one of the most concentrated frenzies of scientific creativity in history. At the same time, he was dealing with personal crises within his family. Letters continued to arrive from his estranged wife that pressed him for money and discussed the guidelines for his contact with their two sons. Through a mutual friend, she demanded that he not ask that his children come visit him in Berlin where they might discover his affair. Einstein assured the friend that in Berlin he was living alone and that his "desolate" apartment had "an almost churchlike atmosphere." The friend replied, referring to Einstein's work on general relativity, "Justifiably so, for unusual divine powers are at work in there."

On the very day that he presented his first paper, he wrote a painfully poignant letter to his elder son, Hans Albert, who was living in Switzerland:

> *Yesterday I received your dear little letter and was delighted with it. I was already afraid you didn't want to write me at all anymore. . . . I shall press for our being together for a month every year so that you see that you have a father who is attached to you and loves you. You can learn a lot of fine and good things from me as well that no one else can offer you so easily. . . . In the last few days I completed*

one of the finest papers of my life; when you are older, I will tell you about it.

He ended with a small apology for seeming so distracted. “I am often so engrossed in my work that I forget to eat lunch,” he wrote.

Einstein also engaged in an awkward interaction with Hilbert. He had been informed that the Göttingen mathematician had spotted the flaws in the *Entwurf* equations. Worried about being scooped, he wrote Hilbert a letter saying that he himself had discovered the flaws, and he sent along a copy of his November 4 lecture.

In his second lecture, delivered on November 11, Einstein imposed new coordinate conditions that allowed his equations to be generally covariant. As it turned out, the change did not greatly improve matters. He was close to the final answer but making little headway. Once again, he sent his paper off to Hilbert and asked him how his own quest was going. “My own curiosity is interfering with my work!” he wrote.

Hilbert sent him a reply that must have unnerved Einstein. He said he had a “solution to your great problem,” and he invited Einstein to come to Göttingen on November 16 and have the dubious pleasure of hearing it. “Since you are so interested, I would like to lay out my theory in very complete detail this coming Tuesday,” Hilbert wrote. “My wife and I would be very pleased if you stayed with us.” Then, after signing his name, Hilbert felt compelled to add a tantalizing and disconcerting postscript. “As far as I understand your new paper, the solution given by you is entirely different from mine.”

Coming to a Head

Einstein wrote four letters on November 15, a Monday, that give a glimpse into his intertwined personal and professional dramas. To Hans Albert, he suggested that he would like to travel to Switzerland at Christmas to visit him. “Maybe it would be better if we were alone somewhere,” such as at a secluded inn, he said to his son. “What do you think?”

He then wrote his estranged wife a conciliatory letter that thanked her for her willingness not “to undermine my relations with the boys.” And he reported to a friend, “I have modified the theory of gravity, having realized

that my earlier proofs had a gap. . . . I shall be glad to come to Switzerland at the turn of the year to see my dear boy."

He also replied to Hilbert and declined his invitation to visit Göttingen the next day. His letter did not hide his anxiety: "The hints you gave in your messages awaken the greatest of expectations. Nevertheless, I must refrain from traveling to Göttingen. . . . I am tired out and plagued by stomach pains. . . . If possible, please send me a correction proof of your study to mitigate my impatience."

As he hurriedly rushed to come up with the precise formulation of his theory, Einstein made a breakthrough that turned his anxiety into elation. He tested a set of revised equations to see if they would yield the correct results for the anomalous shift in Mercury's orbit. The answer came out right: his equations predicted the perihelion should drift by about 43 arc seconds per century. He was so thrilled that he had heart palpitations. "I was beside myself with joy and excitement for days," he told a colleague. To another physicist, he exulted, "The results of Mercury's perihelion movement fill me with great satisfaction. How helpful to us is astronomy's pedantic accuracy, which I used to secretly ridicule!"

The morning of his third lecture, November 18, Einstein received Hilbert's new paper and was dismayed by how similar it was to his own work. His response to Hilbert was terse and clearly designed to assert priority. "The system you furnish agrees—as far as I can see—exactly with what I found in the last few weeks and have presented to the Academy," he wrote. "Today I am presenting to the Academy a paper in which I derive quantitatively out of general relativity, without any guiding hypothesis, the perihelion motion of Mercury. No gravitational theory has achieved this until now."

Hilbert responded kindly and generously the following day, claiming no priority for himself. "Cordial congratulations on conquering perihelion motion," he wrote. "If I could calculate as rapidly as you, in my equations the electron would have to capitulate, and the hydrogen atom would have to produce its note of apology about why it does not radiate." The next day, however, Hilbert sent a paper to a Göttingen science journal describing his own version of the equations for general relativity. The title he picked for his piece was not a modest one: "The Foundations of Physics," he called it.

It is not clear how carefully Einstein read Hilbert's paper or if it affected his thinking as he prepared his climactic fourth lecture at the Prussian Academy. Regardless, he produced in time for his final lecture on November 25—entitled "The Field Equations of Gravitation"—a set of covariant equations that described a general theory of relativity.

It was not nearly as vivid to the layperson as, say, $E = mc^2$. Yet using the condensed notations of tensors, in which sprawling mathematical complexities can be compressed into little subscripts, the crux of the final Einstein field equation is compact enough to be emblazoned on T-shirts worn by physics geeks. In one of its many variations, it can be written as:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = -8 \pi G T_{\mu\nu}$$

The left side of the equation—which is now known as the Einstein tensor and can be written simply as $G_{\mu\nu}$—describes how the geometry of spacetime is warped and curved by massive objects. The right side describes the movement of matter in the gravitational field. The interplay between the two sides shows how objects curve spacetime and how, in turn, this curvature affects the motion of objects.

Both at the time and to this day, there has been a priority dispute over which elements of the mathematical equations of general relativity were discovered first by Hilbert rather than by Einstein. Whatever the case, it was Einstein's theory that was being formalized by these equations, one that he had explained to Hilbert during their time together in Göttingen that summer of 1915. Hilbert graciously noted this in the final version of his paper: "The differential equations of gravitation that result are, as it seems to me, in agreement with the magnificent theory of general relativity established by Einstein." As he later summed it up, "Einstein did the work and not the mathematicians."

Within a few weeks Einstein and Hilbert were repairing their relationship. Hilbert proposed Einstein for membership in the Royal Society of Sciences in Göttingen, and Einstein wrote back with an amiable letter saying how two men who had glimpsed transcendent theories should not be diminished by earthly emotions. "There has been a certain ill-feeling between us, the cause of which I do not want to analyze," Einstein wrote. "I have struggled against the feeling of bitterness attached to it, and this with complete

success. I think of you again with unmixed geniality and ask you to try to do the same with me. Objectively it is a shame when two real fellows who have extricated themselves from this shabby world do not afford each other mutual pleasure."

"The Boldest Dreams"

Einstein's pride was understandable. At age 36, he had produced a dramatic revision of our concept of the universe. His general theory of relativity was not merely the interpretation of some experimental data or the discovery of a more accurate set of laws. It was a whole new way of regarding reality.

With his special theory of relativity, Einstein had shown that space and time did not have independent existences but instead formed a fabric of spacetime. Now, with his general version of the theory, this fabric of spacetime became not merely a container for objects and events. Instead it had its own dynamics that were determined by, and in turn helped to determine, the motion of objects within it—like the way that the fabric of a trampoline will curve as a bowling ball and some billiard balls roll across it and in turn that the dynamic curving of the trampoline fabric will determine the path of the rolling balls and cause the billiard balls to move toward the bowling ball.

The curving and rippling fabric of spacetime explained gravity, its equivalence to acceleration and the general relativity of all forms of motion. In the opinion of Paul Dirac, the Nobel laureate pioneer of quantum mechanics, it was "probably the greatest scientific discovery ever made." And Max Born, another giant of 20th-century physics, called it "the greatest feat of human thinking about nature, the most amazing combination of philosophical penetration, physical intuition and mathematical skill."

The entire process had exhausted Einstein. His marriage had collapsed, and war was ravaging Europe. But he was as happy as he would ever be. "The boldest dreams have now been fulfilled," he exulted to his best friend, engineer Michele Besso. "*General* covariance. Mercury's perihelion motion wonderfully precise." He signed himself "contented but quite worn-out."

Years later, when his younger son, Eduard, asked why he was so famous, Einstein replied by using a simple image to describe his fundamental insight

that gravity was the curving of the fabric of spacetime. “When a blind beetle crawls over the surface of a curved branch, it doesn’t notice that the track it has covered is indeed curved,” he said. “I was lucky enough to notice what the beetle didn’t notice.”

Relativity Primer

General relativity redefined the concept of gravity—rather than a force pulling masses together, the theory exposed it as a simple consequence of the geometry of space and time. The notion grew out of a revelation from the more limited special theory of relativity, which Albert Einstein conceived 10 years earlier. This theory established space and time as a single entity, spacetime (below). In his general theory of relativity, Einstein described what happens when mass is present in spacetime (top right), causing it to curve and forcing objects traveling through it to follow a bent path. If enough mass is packed into a very small region, spacetime becomes infinitely curved, creating a black hole (bottom right).

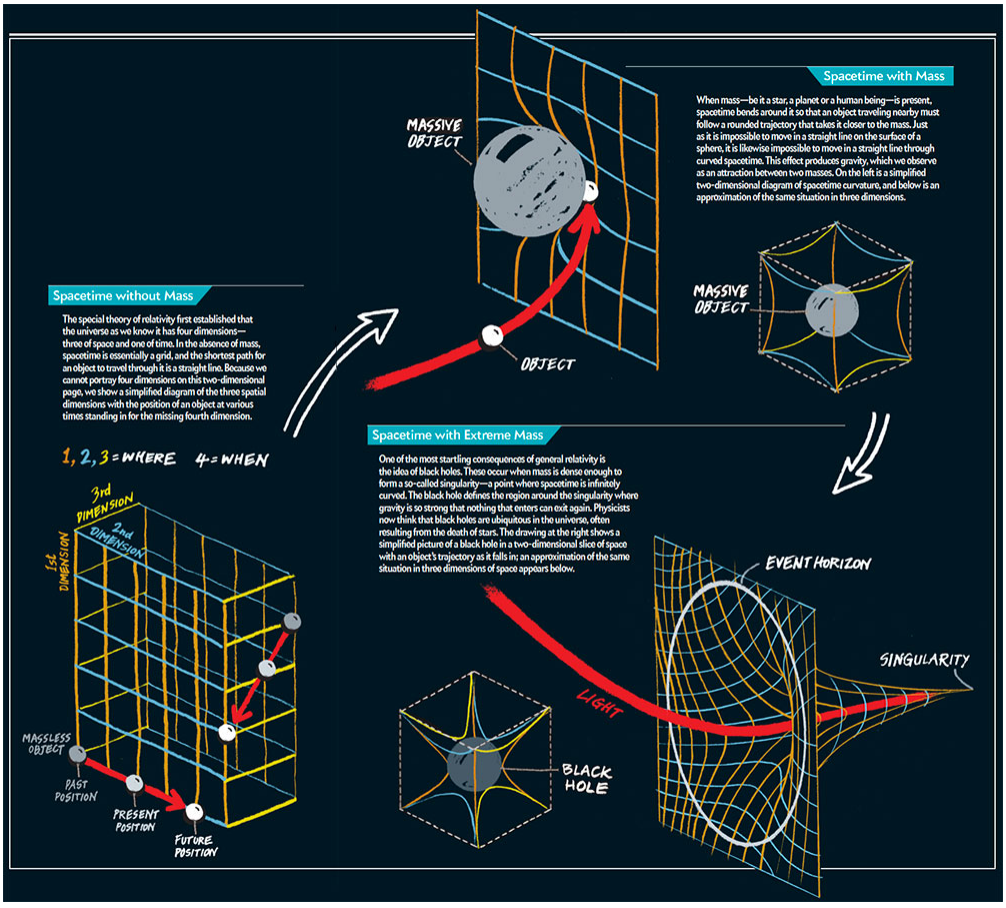


Illustration by Nigel Holmes

--Originally published: Scientific American 313(3); 38-45 (September 2015).

The Strangest Numbers in String Theory

by John C. Baez and John Huerta

As children, we all learn about numbers. We start with counting, followed by addition, subtraction, multiplication and division. But mathematicians know that the number system we study in school is but one of many possibilities. Other kinds of numbers are important for understanding geometry and physics. Among the strangest alternatives is the octonions. Largely neglected since their discovery in 1843, in the past few decades they have assumed a curious importance in string theory. And indeed, if string theory is a correct representation of the universe, they may explain why the universe has the number of dimensions it does.

The Imaginary Made Real

The octonions would not be the first piece of pure mathematics that was later used to enhance our understanding of the cosmos. Nor would it be the first alternative number system that was later shown to have practical uses. To understand why, we first have to look at the simplest case of numbers—the number system we learned about in school—which mathematicians call the real numbers. The set of all real numbers forms a line, so we say that the collection of real numbers is one-dimensional. We could also turn this idea on its head: the line is one-dimensional because specifying a point on it requires one real number.

Before the 1500s the real numbers were the only game in town. Then, during the Renaissance, ambitious mathematicians attempted to solve ever more complex forms of equations, even holding competitions to see who could solve the most difficult problems. The square root of –1 was introduced as a kind of secret weapon by Italian mathematician, physician, gambler and astrologer Gerolamo Cardano. Where others might cavil, he boldly let himself use this mysterious number as part of longer calculations

where the answers were ordinary real numbers. He was not sure why this trick worked; all he knew was that it gave him the right answers. He published his ideas in 1545, thus beginning a controversy that lasted for centuries: Does the square root of –1 really exist, or is it only a trick? Nearly 100 years later no less a thinker than René Descartes rendered his verdict when he gave it the derogatory name “imaginary,” now abbreviated as *i*.

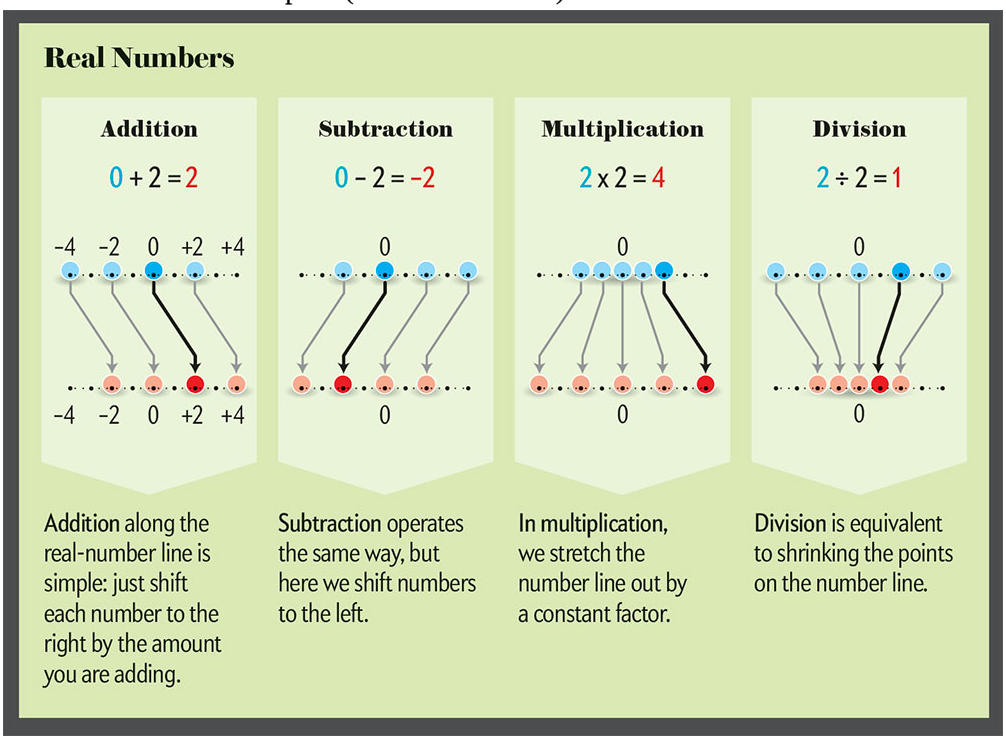
Nevertheless, mathematicians followed in Cardano’s footsteps and began working with complex numbers—numbers of the form $a + bi$, where a and b are ordinary real numbers. Around 1806 Jean-Robert Argand popularized the idea that complex numbers describe points on the plane. How does $a + bi$ describe a point on the plane? Simple: the number a tells us how far left or right the point is, whereas b tells us how far up or down it is.

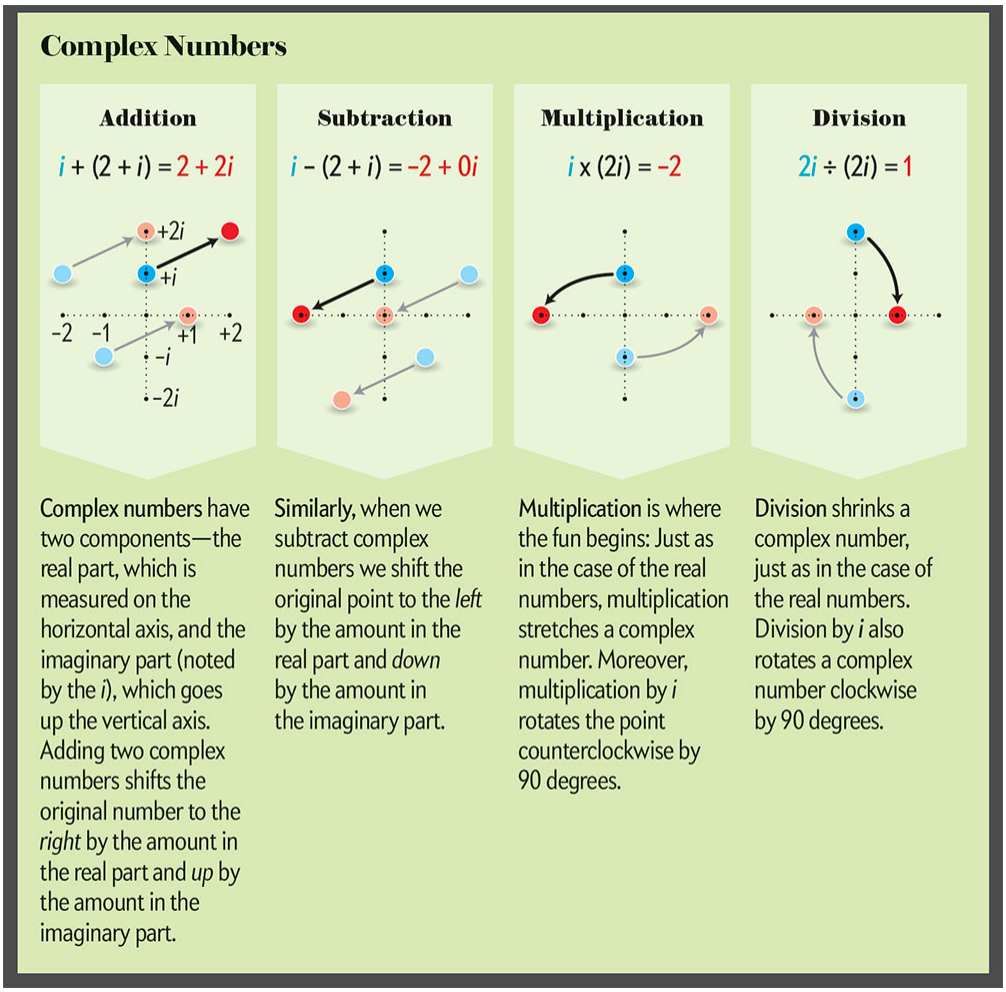
In this way, we can think of any complex number as a point in the plane, but Argand went a step further: he showed how to think of the operations one can do with complex numbers—addition, subtraction, multiplication and division—as geometric manipulations in the plane.

As a warm-up for understanding how these operations can be thought of as geometric manipulations, first think about the real numbers. Adding or subtracting any real number slides the real line to the right or left. Multiplying or dividing by any positive number stretches or squashes the line. For example, multiplying by 2 stretches the line by a factor of 2, whereas dividing by 2 squashes it down, moving all the points twice as close as they were. Multiplying by –1 flips the line over.

The same procedure works for complex numbers, with just a few extra twists. Adding any complex number $a + bi$ to a point in the plane slides that point right (or left) by an amount a and up (or down) by an amount b. Multiplying by a complex number stretches or squashes but also rotates the complex plane. In particular, multiplying by i rotates the plane a quarter turn. Thus, if we multiply 1 by i twice, we rotate the plane a full half-turn from the starting point to arrive at –1. Division is the opposite of multiplication, so to divide we just shrink instead of stretching, or vice versa, and then rotate in the opposite direction.

In grade school we are taught to connect the abstract ideas of addition and subtraction to concrete operations—moving numbers up and down the number line. This connection between algebra and geometry turns out to be incredibly powerful. Because of it, mathematicians can use the algebra of the octonions to solve problems in hard-to-imagine eight-dimensional worlds. The panels below show how to extend algebraic operations on the real-number line to complex (two-dimensional) numbers.





Illustrations by Brown Bird Design

Almost everything we can do with real numbers can also be done with complex numbers. In fact, most things work better, as Cardano knew, because we can solve more equations with complex numbers than with real numbers. But if a two-dimensional number system gives the user added calculating power, what about even higher-dimensional systems? Unfortunately, a simple extension turns out to be impossible. An Irish mathematician would uncover the secret to higher-dimensional number systems decades later. And only now, two centuries on, are we beginning to understand how powerful they can be.

Hamilton's Alchemy

In 1835, at the age of 30, mathematician and physicist William Rowan Hamilton discovered how to treat complex numbers as pairs of real numbers. At the time mathematicians commonly wrote complex numbers in

the form $a + bi$ that Argand popularized, but Hamilton noted that we are also free to think of the number $a + bi$ as just a peculiar way of writing two real numbers—for instance (a, b).

This notation makes it very easy to add and subtract complex numbers—just add or subtract the corresponding real numbers in the pair. Hamilton also came up with slightly more involved rules for how to multiply and divide complex numbers so that they maintained the nice geometric meaning discovered by Argand.

After Hamilton invented this algebraic system for complex numbers that had a geometric meaning, he tried for many years to invent a bigger algebra of triplets that would play a similar role in three-dimensional geometry, an effort that gave him no end of frustrations. He once wrote to his son, “Every morning ... on my coming down to breakfast, your (then) little brother William Edwin, and yourself, used to ask me: ‘Well, Papa, can you multiply triplets?’ Whereto I was always obliged to reply, with a sad shake of the head: ‘No, I can only add and subtract them.’ ” Although he could not have known it at the time, the task he had given himself was mathematically impossible.

Hamilton was searching for a three-dimensional number system in which he could add, subtract, multiply and divide. Division is the hard part: a number system where we can divide is called a division algebra. Not until 1958 did three mathematicians prove an amazing fact that had been suspected for decades: any division algebra must have dimension one (which is just the real numbers), two (the complex numbers), four or eight. To succeed, Hamilton had to change the rules of the game.

Hamilton himself figured out a solution on October 16, 1843. He was walking with his wife along the Royal Canal to a meeting of the Royal Irish Academy in Dublin when he had a sudden revelation. In three dimensions, rotations, stretching and shrinking could not be described with just three numbers. He needed a fourth number, thereby generating a four-dimensional set called quaternions that take the form $a + bi + cj + dk$. Here the numbers i, j and k are three different square roots of $–1$.

Hamilton would later write: “I then and there felt the galvanic circuit of thought close; and the sparks which fell from it were the fundamental

equations between *i*, *j* and *k*; exactly such as I have used them ever since." And in a noteworthy act of mathematical vandalism, he carved these equations into the stone of the Brougham Bridge. Although they are now buried under graffiti, a plaque has been placed there to commemorate the discovery.

It may seem odd that we need points in a four-dimensional space to describe changes in three-dimensional space, but it is true. Three of the numbers come from describing rotations, which we can see most readily if we imagine trying to fly an airplane. To orient the plane, we need to control the pitch, or angle with the horizontal. We also may need to adjust the yaw, by turning left or right, as a car does. And finally, we may need to adjust the roll: the angle of the plane's wings. The fourth number we need is used to describe stretching or shrinking.

Hamilton spent the rest of his life obsessed with the quaternions and found many practical uses for them. Today in many of these applications the quaternions have been replaced by their simpler cousins: vectors, which can be thought of as quaternions of the special form $ai + bj + ck$ (the first number is just zero). Yet quaternions still have their niche: they provide an efficient way to represent three-dimensional rotations on a computer and show up wherever this is needed, from the attitude-control system of a spacecraft to the graphics engine of a video game.

Imaginaries without End

Despite these applications, we might wonder what, exactly, are *j* and *k* if we have already defined the square root of –1 as *i*. Do these square roots of –1 really exist? Can we just keep inventing new square roots of –1 to our heart's content?

These questions were asked by Hamilton's college friend, a lawyer named John Graves, whose amateur interest in algebra got Hamilton thinking about complex numbers and triplets in the first place. The very day after his fateful walk in the fall of 1843, Hamilton sent Graves a letter describing his breakthrough. Graves replied nine days later, complimenting Hamilton on the boldness of the idea but adding, "There is still something in the system which gravels me. I have not yet any clear views as to the extent to which we are at liberty arbitrarily to create imaginaries, and to endow them with

supernatural properties.” And he asked: “If with your alchemy you can make three pounds of gold, why should you stop there?”

Like Cardano before him, Graves set his concerns aside for long enough to conjure some gold of his own. On December 26 he wrote again to Hamilton, describing a new eight-dimensional number system that he called the octaves and that are now called octonions. Graves was unable to get Hamilton interested in his ideas, however. Hamilton promised to speak about Graves’s octaves at the Irish Royal Society, which is one way mathematical results were published at the time. But Hamilton kept putting it off, and in 1845 the young genius Arthur Cayley rediscovered the octonions and beat Graves to publication. For this reason, the octonions are also sometimes known as Cayley numbers.

Why didn’t Hamilton like the octonions? For one thing, he was obsessed with research on his own discovery, the quaternions. He also had a purely mathematical reason: the octonions break some cherished laws of arithmetic.

The quaternions were already a bit strange. When you multiply real numbers, it does not matter in which order you do it—2 times 3 equals 3 times 2, for example. We say that multiplication commutes. The same holds for complex numbers. But quaternions are noncommutative. The order of multiplication matters.

Order is important because quaternions describe rotations in three dimensions, and for such rotations the order makes a difference to the outcome. You can check this out yourself. Take a book, flip it top to bottom (so that you are now viewing the back cover) and give it a quarter turn clockwise (as viewed from above). Now do these two operations in reverse order: first rotate a quarter turn, then flip. The final position has changed. Because the result depends on the order, rotations do not commute.

The Problem with Rotations

Ordinarily you can multiply numbers together in whatever order you like. For example, 2 times 3 is the same as 3 times 2. In higher-dimensional number systems such as the quaternions and octonions, however, order is very important. Consider the quaternions, which describe rotations in three dimensions. If we take an object such as a book, the order in which we rotate it has a great effect on its final orientation. In the top row at the right, we flip the book vertically, then rotate it, revealing the page edges. In the bottom row, rotating the book and then flipping reveal the spine on the opposite side.

The octonions are much stranger. Not only are they noncommutative, they also break another familiar law of arithmetic: the associative law $(xy)z = x(yz)$. We have all seen a nonassociative operation in our study of mathematics: subtraction. For example, (3 – 2) – 1 is different from 3 – (2 – 1). But we are used to multiplication being associative, and most mathematicians still feel this way, even though they have gotten used to noncommutative operations. Rotations are associative, for example, even though they do not commute.

But perhaps most important, it was not clear in Hamilton's time just what the octonions would be good for. They are closely related to the geometry of seven and eight dimensions, and we can describe rotations in those dimensions using the multiplication of octonions. But for more than a century that was a purely intellectual exercise. It would take the development of modern particle physics—and string theory in particular—to see how the octonions might be useful in the real world.

Symmetry and Strings

In the 1970s and 1980s theoretical physicists developed a strikingly beautiful idea called supersymmetry. (Later researchers would learn that string theory requires supersymmetry.) It states that at the most fundamental levels, the universe exhibits a symmetry between matter and the forces of nature. Every matter particle (such as an electron) has a partner particle that

carries a force. And every force particle (such as a photon, the carrier of the electromagnetic force) has a twin matter particle.

Supersymmetry also encompasses the idea that the laws of physics would remain unchanged if we exchanged all the matter and force particles. Imagine viewing the universe in a strange mirror that, rather than interchanging left and right, traded every force particle for a matter particle, and vice versa. If supersymmetry is true, if it truly describes our universe, this mirror universe would act the same as ours. Even though physicists have not yet found any concrete experimental evidence in support of supersymmetry, the theory is so seductively beautiful and has led to so much enchanting mathematics that many physicists hope and expect that it is real.

One thing we know to be true, however, is quantum mechanics. And according to quantum mechanics, particles are also waves. In the standard three-dimensional version of quantum mechanics that physicists use every day, one type of number (called spinors) describes the wave motion of matter particles. Another type of number (called vectors) describes the wave motion of force particles. If we want to understand particle interactions, we have to combine these two using a cobbled-together simulacrum of multiplication. Although the system we use right now might work, it is not very elegant at all.

As an alternative, imagine a strange universe with no time, only space. If this universe has dimension one, two, four or eight, both matter and force particles would be waves described by a single type of number—namely, a number in a division algebra, the only type of system that allows for addition, subtraction, multiplication and division. In other words, in these dimensions the vectors and spinors coincide: they are each just real numbers, complex numbers, quaternions or octonions, respectively. Supersymmetry emerges naturally, providing a unified description of matter and forces. Simple multiplication describes interactions, and all particles—no matter the type—use the same number system.

Yet our plaything universe cannot be real, because we need to take time into account. In string theory, this consideration has an intriguing effect. At any moment in time a string is a one-dimensional thing, like a curve or line. But this string traces out a two-dimensional surface as time passes. This

evolution changes the dimensions in which supersymmetry arises, by adding two—one for the string and one for time. Instead of supersymmetry in dimension one, two, four or eight, we get supersymmetry in dimension three, four, six or 10.

Coincidentally string theorists have for years been saying that only 10-dimensional versions of the theory are self-consistent. The rest suffer from glitches called anomalies, where computing the same thing in two different ways gives different answers. In anything other than 10 dimensions, string theory breaks down. But 10-dimensional string theory is, as we have just seen, the version of the theory that uses octonions. So if string theory is right, the octonions are not a useless curiosity: on the contrary, they provide the deep reason why the universe must have 10 dimensions: in 10 dimensions, matter and force particles are embodied in the same type of numbers—the octonions.

But this is not the end of the story. Recently physicists have started to go beyond strings to consider membranes. For example, a two-dimensional membrane, or 2-brane, looks like a sheet at any instant. As time passes, it traces out a three-dimensional volume in spacetime.

Whereas in string theory we had to add two dimensions to our standard collection of one, two, four and eight, now we must add three. Thus, when we are dealing with membranes we would expect supersymmetry to naturally emerge in dimensions four, five, seven and 11. And as in string theory we have a surprise in store: researchers tell us that M-theory (the "M" typically stands for "membrane") requires 11 dimensions—implying that it should naturally make use of octonions. Alas, nobody understands M-theory well enough to even write down its basic equations (that M can also stand for "mysterious"). It is hard to tell precisely what shape it might take in the future.

At this point we should emphasize that string theory and M-theory have as of yet made no experimentally testable predictions. They are beautiful dreams—but so far only dreams. The universe we live in does not look 10- or 11-dimensional, and we have not seen any symmetry between matter and force particles. David Gross, one of the world's leading experts on string theory, currently puts the odds of seeing some evidence for supersymmetry

at CERN’s Large Hadron Collider at 50 percent. Skeptics say they are much less. Only time will tell.

Because of this uncertainty, we are still a long way from knowing if the strange octonions are of fundamental importance in understanding the world we see around us or merely a piece of beautiful mathematics. Of course, mathematical beauty is a worthy end in itself, but it would be even more delightful if the octonions turned out to be built into the fabric of nature. As the story of the complex numbers and countless other mathematical developments demonstrates, it would hardly be the first time that purely mathematical inventions later provided precisely the tools that physicists need.

--Originally published: Scientific American 304(5); 60-65 (May 2011).

Walls of Water

by Dana Mackenzie

All along the gulf of Mexico, 2010 was the summer of the Oil Spill. As BP's uncapped Deepwater Horizon oil well gushed away off of Louisiana, tourists stayed away from the Gulf Coast in droves, convinced by news reports that oil was coming ashore or would do so imminently. As far away as Fort Myers and Key Largo in Florida, beaches were deserted and hotel occupancy rates were down.

In reality, the situation was never so dire—especially on the western coast of Florida. This part of the Gulf Coast was protected for the duration of the oil spill by a persistent, invisible divide. Lying above the continental shelf off of Florida was an unseen line that directed the oil and prevented it from spreading farther east. It was not a solid object, but a wall of water that moved around as ocean currents shifted. Nevertheless, this wall was just as effective as any seawall or containment boom.

Scientists call these invisible walls "transport barriers," and they are the maritime equivalent of continental divides. They separate water flowing in one direction from water flowing in another. In a chaotic ocean, they provide a road map to tell you where the traffic is going. Although water currents often appear to be almost completely unpredictable, transport barriers restore a measure of order and structure to their chaotic flow.

The study of these structures has blossomed in recent years, and their importance is still not fully appreciated by the scientific community. But already researchers have shown how their study may help explain why the surface oil from the Gulf spill disappeared more rapidly than expected and why none of it escaped through the Strait of Florida into the Atlantic. During future disasters, understanding these flows could make cleanup efforts more efficient. The research could also elucidate how blood flow

affects the formation of plaques in arteries and help to predict how allergy-causing spores migrate in the atmosphere.

The study of chaos came of age in the 1970s, when scientists discovered that in certain natural phenomena, even tiny perturbations could lead to profound changes. The proverbial refrain is that the flutter of a butterfly's wing on one side of the globe could make subtle changes in air currents that cascaded, to the point of causing a tornado on the other side weeks later.

Flowing fluids—which include gases such as air and liquids such as seawater—are in fact the quintessential example of chaotic systems and one of the most ubiquitous: the dynamics of fluids govern phenomena from the Gulf Stream to the flow of air through a wind turbine to curving penalty kicks in soccer. The mathematical equations describing fluid flow were unveiled nearly 200 years ago, by Claude-Louis Navier (in 1822) and George Stokes (in 1842). Yet knowing the equations is not the same thing as solving them, and the Navier-Stokes equations remain among the most challenging problems in mathematics.

In principle, an exact solution of the Navier-Stokes equations would yield a detailed prediction of the future behavior of a fluid. But the precision of the answer would depend on exact knowledge of the present—or what scientists call the initial conditions. In reality, you can never know where every molecule of water in the ocean is going, and in a chaotic system any uncertainties—like the effects of a butterfly's motions—grow exponentially over time. Your exact solution to the Navier- Stokes equations will rapidly become moot.

And yet "chaotic" does not mean "random" or "unpredictable," at least in principle. In the past decade or so mathematicians have created a theoretical framework for understanding the persistent structures such as transport barriers that are hidden in chaotic fluids. In 2001 George Haller, a mathematician now at McGill University, gave these structures the rather unwieldy name "Lagrangian coherent structures." More poetically, Haller calls the intricate structure of transport barriers "the skeleton of turbulence." Once you have identified these structures in a body of fluid, you can make useful short- to medium-term predictions of where the fluid flow will carry an object, for instance, even without a perfect, precise solution of the Navier-Stokes equations.

What does a transport barrier look like? You are looking at one every time you see a smoke ring. At its core lies an *attracting* Lagrangian coherent structure—a curve toward which particles flow, as if they were attracted by a magnet. Ordinarily you cannot see such a structure, but if you blow smoke into the air, the smoke particles will concentrate around it and make it visible.

Much harder to visualize are the *repelling* Lagrangian coherent structures —curves that, if they were visible, would appear as if they were pushing particles away. If you could run time backward, they would be easier to see (because they would attract particles); failing that, the only way to find them is to tease them out by computer analysis. Though difficult to observe, repelling structures are particularly important because, as Haller has proved mathematically, they tend to form transport barriers.

An experiment conducted in the summer of 2003 in Monterey Bay off the coast of California showed that Lagrangian coherent structures could be computed in real time and in real bodies of water. Mathematician Shawn C. Shadden of the Illinois Institute of Technology and his collaborators monitored surface currents in the bay using four high-frequency radar stations deployed around the bay.

Analyzing the radar data, the researchers discovered that most of the time a long transport barrier snakes across the bay from Point Pinos, at the southern edge, almost all the way to the northern side. Waters to the east of the barrier circulate back into the bay, whereas those on the western side go out to sea. (Occasionally the barrier detaches from Point Pinos and drifts farther out to sea.) Such information could be vital in case of a pollutant spill.

To confirm that the computed structures did actually behave as advertised, Shadden’s team tracked the motion of four drifting buoys they deployed in collaboration with the Monterey Bay Aquarium Research Institute. When they placed drifters on opposite sides of the transport barrier, one drifter would follow the water circulating back into the bay, and the other one would hitch a ride on the currents heading southward along the coast. They also showed that a drifter placed on the recirculating side of the structure would stay in the bay for 16 days—even though they had used only three days of data to compute it. This robustness of their results testified to the

strength and persistence of the transport barrier. For 16 days, it really was like an invisible wall in the water.

Close Call in the Gulf

The most dramatic demonstration of the transport barrier concept came in the aftermath of the 2010 Gulf oil spill. Oceanographers and mathematicians have analyzed the huge volumes of data on the spill and shown how the information could have enabled scientists to better predict where the oil would go.

Lagrangian coherent structures might help explain why the surface oil disappeared more rapidly than anyone expected—much faster, for example, than the oil from the *Exxon Valdez* spill in 1989 in Prince William Sound in Alaska. (The fate of the subsurface oil has been more controversial, and much of it may still remain at the bottom of the Gulf.) The warm Gulf of Mexico, it turned out, is home to hordes of microorganisms that feed on hydrocarbons that naturally seep into the Gulf waters. Given a much more abundant supply of hydrocarbons than usual, these microorganisms flourished. Microbiologist Dave Valentine and mathematician Igor Mezic, both at the University of California, Santa Barbara, showed that the bacteria tended to congregate in coherent regions defined by transport barriers. Clearly, the long-term stability of those regions helped the oil degrade. Valentine notes that it would have been a different story if the blowout had happened off the coast of Brazil, another region where vast deepwater oil reserves have been discovered. There the currents lead out to sea, where a captive supply of bacteria does not exist to chow down on the hydrocarbons.

Transport barriers may also explain why the oil from Deepwater Horizon avoided flowing into the Loop Current, a persistent jet that leads through the Florida Straits and into the Atlantic, where it could have polluted beaches along the East Coast. As late as July 2, the National Oceanic and Atmospheric Administration was predicting a 61 to 80 percent chance some oil would make it to the Loop Current. The prediction was based on 15 years of historical ocean current data from the Gulf of Mexico.

In 2010 we apparently got lucky. First, unusually strong winds from the Southwest pushed the oil slick to the north, away from the Loop Current. In

addition, a giant eddy, called Eddy Franklin, detached from the Loop Current and pushed it farther south than usual, forming a barrier between the oil and the current. It remains to be seen whether any of these phenomena could have been anticipated. Haller, however, with oceanographer Maria Olascoaga of the University of Miami, has shown that other seemingly capricious changes in the oil slick were predictable. On May 17, for instance, a giant “tiger tail” (named after its shape) of oil suddenly traveled more than 160 kilometers southeast in one day. According to their computer analysis, the tiger tail traveled along an attracting Lagrangian coherent structure, and the impending instability was presaged seven days earlier by the formation of a strong attracting “core” on that structure. Likewise, an abrupt westward retreat of the oil slick’s leading edge on June 16 was anticipated nine days earlier by the formation of an exceptionally strong repelling core to the east of the slick. Had surveillance been in place that could identify transport barriers, cleanup boats could have been sent to the right locations.

Beyond the study of oceanic currents, applications of the transport barrier concept have proliferated in recent years. For example, Shane Ross of Virginia Polytechnic Institute has studied the effect of transport barriers in the atmosphere on airborne pathogens. He and plant biologist David Schmale, also at Virginia Tech, used a small drone airplane to collect air samples at an altitude between tens and hundreds of meters above Blacksburg. When an attracting structure passed by or when two repelling structures passed in rapid succession, the researchers detected a spike in the number of spores of a fungus called *Fusarium*. Ross hypothesizes that in the first case the spores had been pulled toward the coherent structures, whereas in the second they had become trapped between the two repelling barriers, like cattle herded into a small region by prods. Some of the spores were of a species that does not usually occur in Virginia, which suggests that the structures remained intact long enough for the spores to be transported several hundred kilometers.

Shadden is now studying the role of Lagrangian coherent structures in blood flow. For example, he has used these structures to reveal the boundaries between blood ejected on one heartbeat and blood ejected on the next. He showed that most of the blood in a normal ventricle remains there for at most two heartbeats. But in six patients with enlarged hearts, regions

of blood recirculated for much longer—“a widely recognized risk factor for thrombosis,” he wrote in a draft of his study.

More than a decade after Haller named them, Lagrangian coherent structures are still far from being a mainstream tool in oceanography or atmospheric science. One objection raised about their usefulness is that if there are errors in the measurement of the flow field, they will surely propagate and produce errors in the predictions of the transport barrier as well. But the Monterey Bay experiment found that the location of the transport barriers was relatively insensitive to measurement errors.

Another objection is that to compute the structures, you need to know the entire flow field, meaning the velocity of water flowing at each point. But if you know that, you can forecast the oil slick using existing computer models. So what are calculations of Lagrangian coherent structures good for?

As it turns out, forecasting is not the only game in town. “Hindcasting” may turn out to be important in finding the source of “mystery oil spills” that wash ashore from unknown sources—often from sunken ships. For example, the *SS Jacob Luckenbach*, which sank off San Francisco in 1953, polluted the California coast every year beginning around 1991, but the source of the spill was not discovered until 2002. Plane crashes and shipwrecks have also produced “debris spills” and “body spills.” Because conventional ocean models cannot be reversed in time, rescuers cannot extrapolate backward from the ob - served debris field to find the source. Oceanographer C. J. Beegle-Krause and mathematician Thomas Peacock of the Massachusetts Institute of Technology are now working on using La grangian coherent structures to forecast where shipwreck survivors will drift in the currents, which would help narrow down the search area. In such situations, as Peacock notes, “even a few minutes might be a matter of life and death.”

Finally, Lagrangian coherent structures provide more than mere forecasts or hindcasts; they provide understanding. Knowing about the structures enables scientists to better interpret the predictions of computer models. If a model predicts that a filament of oil will move toward Pensacola and we can see a structure pushing it or pulling it that way, we can be reasonably

confident in the prediction. If there is no corresponding structure, we might treat the model with more skepticism.

Mathematicians are now broadening their research into different types of organized structures in turbulent fluids, such as eddies and jets. With deeper understanding, we may be able to answer questions about chaotic phenomena that now elude us.

--Originally published: Scientific American 309(1); 86-89 (July 2013).

The Particle Code

by Matthew von Hippel

The Large Hadron Collider, or LHC, is the biggest machine humans have ever built. Pooling the resources of more than 100 countries, it accelerates protons to within a millionth of a percent of the speed of light. When they collide, the protons break into their component parts (quarks and the gluon particles that glue them together) and create particles that were not there before. This is how, in 2012, the LHC achieved the first detection of a Higgs boson, the final missing particle predicted by the Standard Model of particle physics. Now physicists hope the LHC will find something genuinely new: particles not already in their current theory—particles that explain the mystery of dark matter, for instance, or offer solutions to other lingering questions. For such a discovery, scientists must pore through the 30 petabytes a year of data the machine produces to identify tiny deviations where the results do not quite match the Standard Model.

Of course, all of that effort will be useless if we do not know what the Standard Model predicts.

That is where I come in. The questions we want to ask about the LHC come in the form of probabilities. “What is the chance that two protons bounce off each other?” “How often will we produce a Higgs boson?” Scientists compute these probabilities with “scattering amplitudes,” formulas that tell us how likely it is that particles “scatter” (essentially, bounce) off each other in a particular way. I am part of a group of physicists and mathematicians who work to speed up these calculations and find better tricks than the old, cumbersome methods handed down by our scientific forebears. We call ourselves “amplitudeologists.”

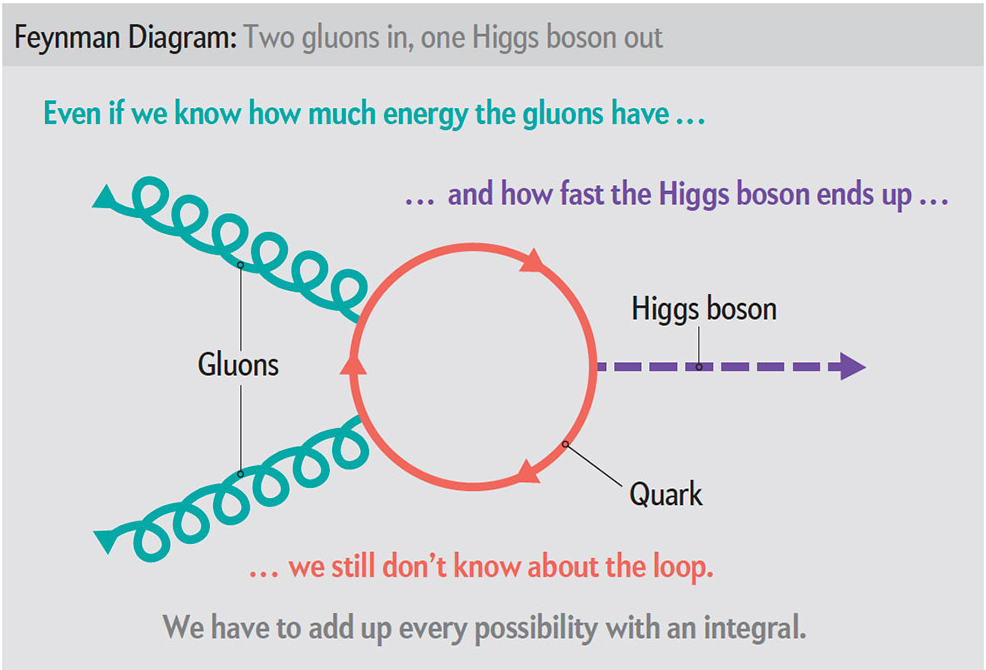
Amplitudeologists trace our field back to the research of two physicists, Stephen Parke and Tomasz Taylor. In 1986 they found a single formula that

described collisions between any number of gluons, simplifying what would ordinarily be pages of careful case-by-case calculations. The field actually kicked off in the 1990s and early 2000s, when a slew of new methods promised to streamline a wide variety of particle physics computations. Nowadays amplitudeology is booming: the Amplitudes 2018 conference had 160 participants, and 100 attended the summer school the week before, aimed at training young researchers in the tricks of the field. We have gotten some public attention, too: physicists Nima Arkani-Hamed and Jaroslav Trnka's Amplituhedron (a way to describe certain amplitudes in the language of geometry) made the news in 2013, and on television *The Big Bang Theory's* Sheldon Cooper has been known to dabble in amplitudeology.

Lately we have taken a big step forward, moving beyond the basic tools we have already developed into more complex techniques. We are entering a realm of calculations sensitive enough to match the increasing precision of the LHC. With these new tools we stand ready to detect even tiny differences between Standard Model predictions and the reality inside the LHC, potentially allowing us to finally reveal the undiscovered particles physicists dream of.

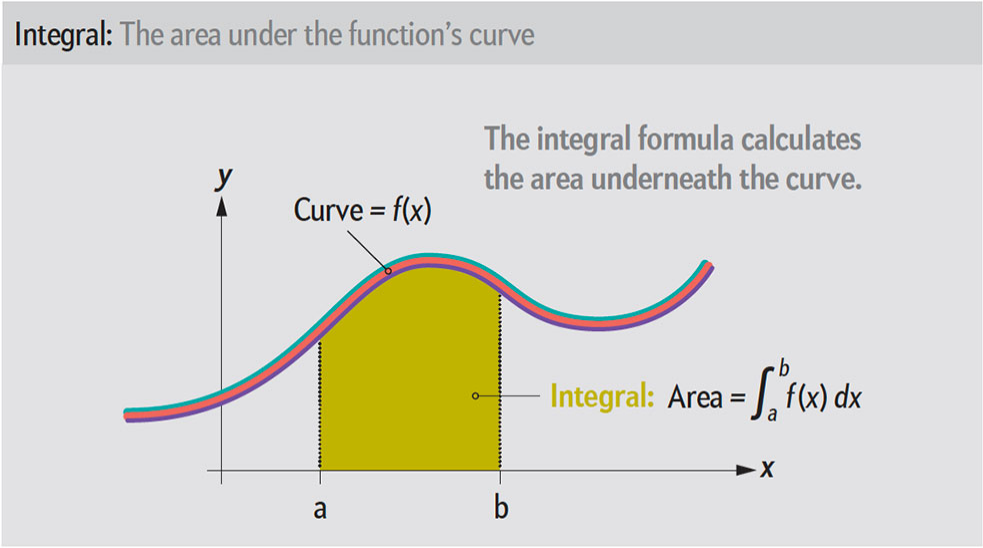
Loops and Lines

To organize our calculations, scientists have long used pictures called Feynman diagrams. Invented by physicist Richard Feynman in 1948, these figures depict paths along which particles travel. Suppose we want to know the chance that two gluons merge and form a Higgs boson. We start by drawing lines representing the particles we know about: two gluons going in and one Higgs boson coming out. We then have to connect those lines by drawing more particle lines in the middle of the diagram, according to the rules of the Standard Model. These additional particles may be "virtual": that is, they are not literally particles in the way the gluons and Higgs are in our picture. Instead they are shorthand, a way to keep track of how different quantum fields can interact.



Credit: Illustration by Jen Christiansen.

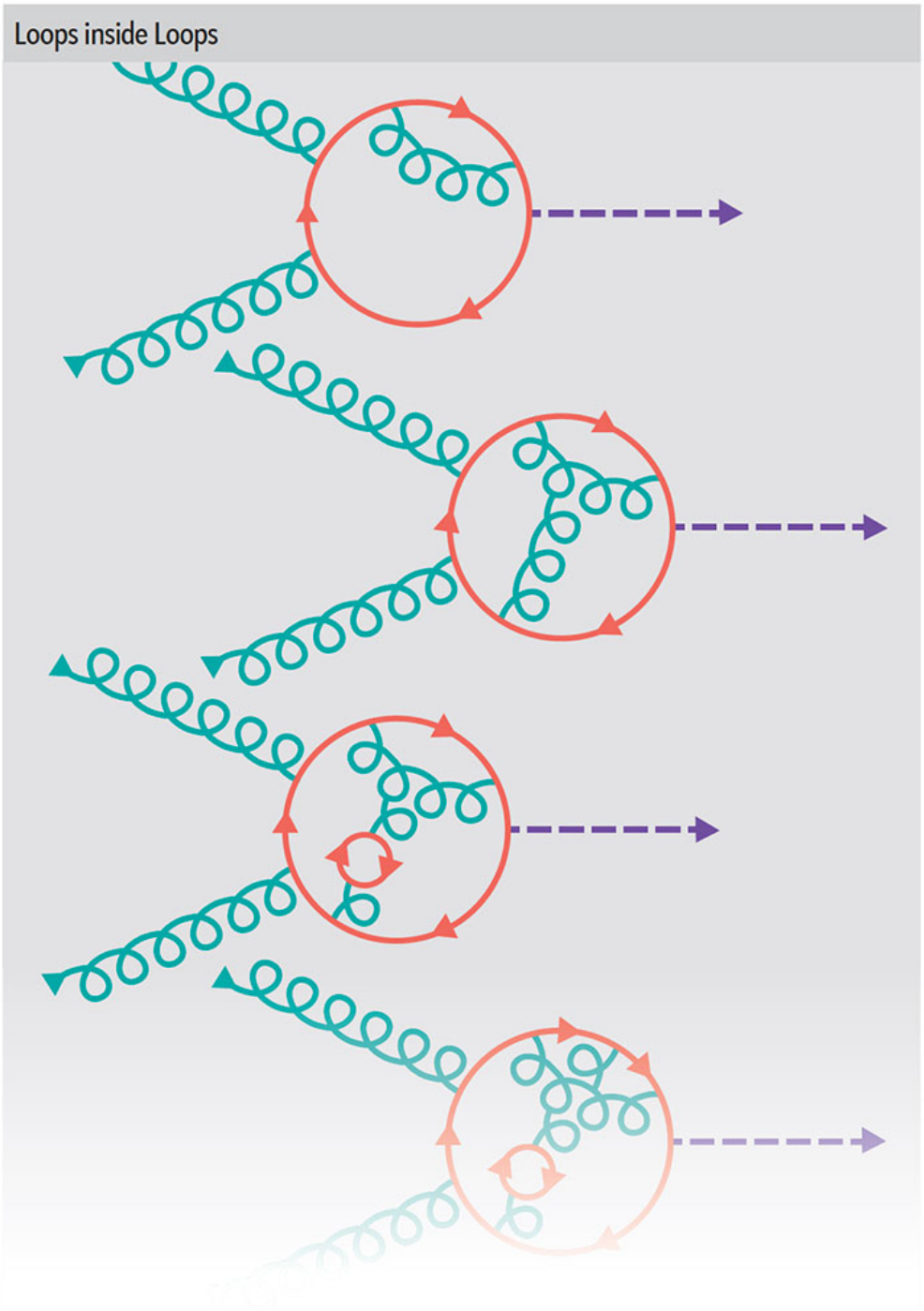
Feynman diagrams are not just pretty pictures—they are instructions, telling us to use information about the particles we draw to calculate a probability. If we know the speed and energy of the gluons and Higgs boson in our diagram, we can try to work out the properties of the virtual particles in between. Sometimes, though, the answer is uncertain. Trace your finger along the particle paths, and you might find a closed loop: a path that ends up back where you started. A particle traveling in a loop like that is not "input" or "output": its properties never get measured. We do not know how fast it is going or how much energy it has. Though counterintuitive, it is a consequence of the fundamental uncertainty of quantum mechanics, which prevents us from measuring two traits of a particle, such as speed and position, at the same time. Quantum mechanics tells us how to deal with this uncertainty—we have to add up every possibility, summing the probabilities for any speed and energy the virtual particles could have, using a technique you might remember from high school calculus: an integral.



Credit: Illustration by Jen Christiansen.

In principle, to calculate a scattering amplitude we have to draw every diagram that could possibly connect our particles, every way the starting ingredients could have turned into the finished products (here the pair of gluons and the Higgs boson). That is a lot of diagrams, an infinite number, in fact: we could keep drawing loops inside loops as far as we like, requiring us to calculate more and more complicated integrals each time.

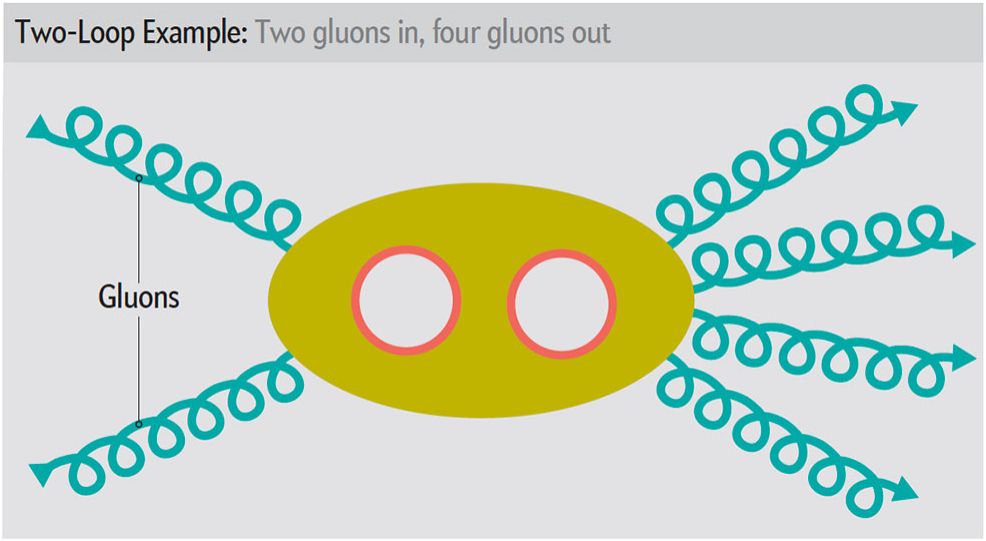
In practice, we are saved by the low strength of most quantum forces. When a group of lines in a diagram connect, it depicts an "interaction" among different types of particles. Each time this happens we have to multiply by a constant, related to the strength of the force that makes the particles interact. If we want to draw a diagram with more closed loops, we have to connect up more lines and multiply by more of these constants. For electricity and magnetism, the relevant constants are small: for each loop you add, you divide by roughly 137. This means that the diagrams with more and more loops make up a smaller and smaller piece of your final answer, and eventually that piece is so small that the experiments cannot detect it. The most careful experiments on electricity and magnetism are accurate up to an astounding 10 decimal places, some of the most precise measurements in all of science. Getting that far requires "only" four loops, four factors off $^1/_{137}$ before the number you are calculating is too small to measure. In many cases, these numbers have actually been calculated, and all 10 decimal places agree with experiments.



Credit: Illustration by Jen Christiansen.

The strong nuclear force is a tougher beast. It is the force that glues together protons and neutrons and the quarks inside them. It is quite a bit stronger than electricity and magnetism: for calculations at the LHC, each loop means dividing not by 137 but by 10. Getting up to 10 digits of precision would mean drawing 10 loops.

The LHC is not as precise as those electricity and magnetism experiments. At the moment, measurements from the machine are just starting to match the precision of two-loop calculations. Still, those results are already quite messy. For example, a twoloop calculation in 2010 by physicists Vittorio Del Duca, Claude Duhr and Vladimir Smirnov computed the chance that two gluons collide and four gluons come out. They made their calculation using a simplified theory, with some special shortcuts, and the resulting formula still clocked in at 17 pages of complicated integrals. That length was not too surprising; everyone knew that two-loop calculations were hard.



Credit: Illustration by Jen Christiansen.

Until a few months later, when another group managed to write the same result on two lines. That group was a collaboration among three physicists —Marcus Spradlin, Cristian Vergu and Anastasia Volovich—and a mathematician, Alexander B. Goncharov. The trick they used was extraordinarily powerful, and it exposed amplitudeologists to an area of mathematics that most of us had not seen before, one that has driven my career to this day.

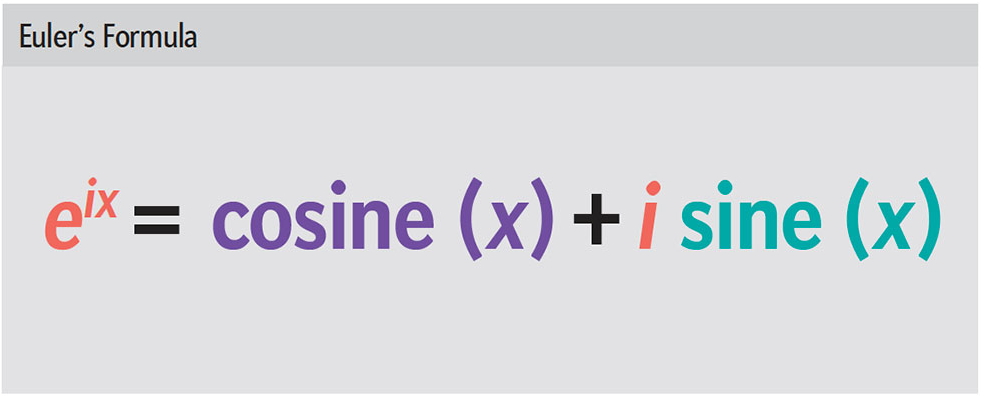
PERIODS AND LOGS

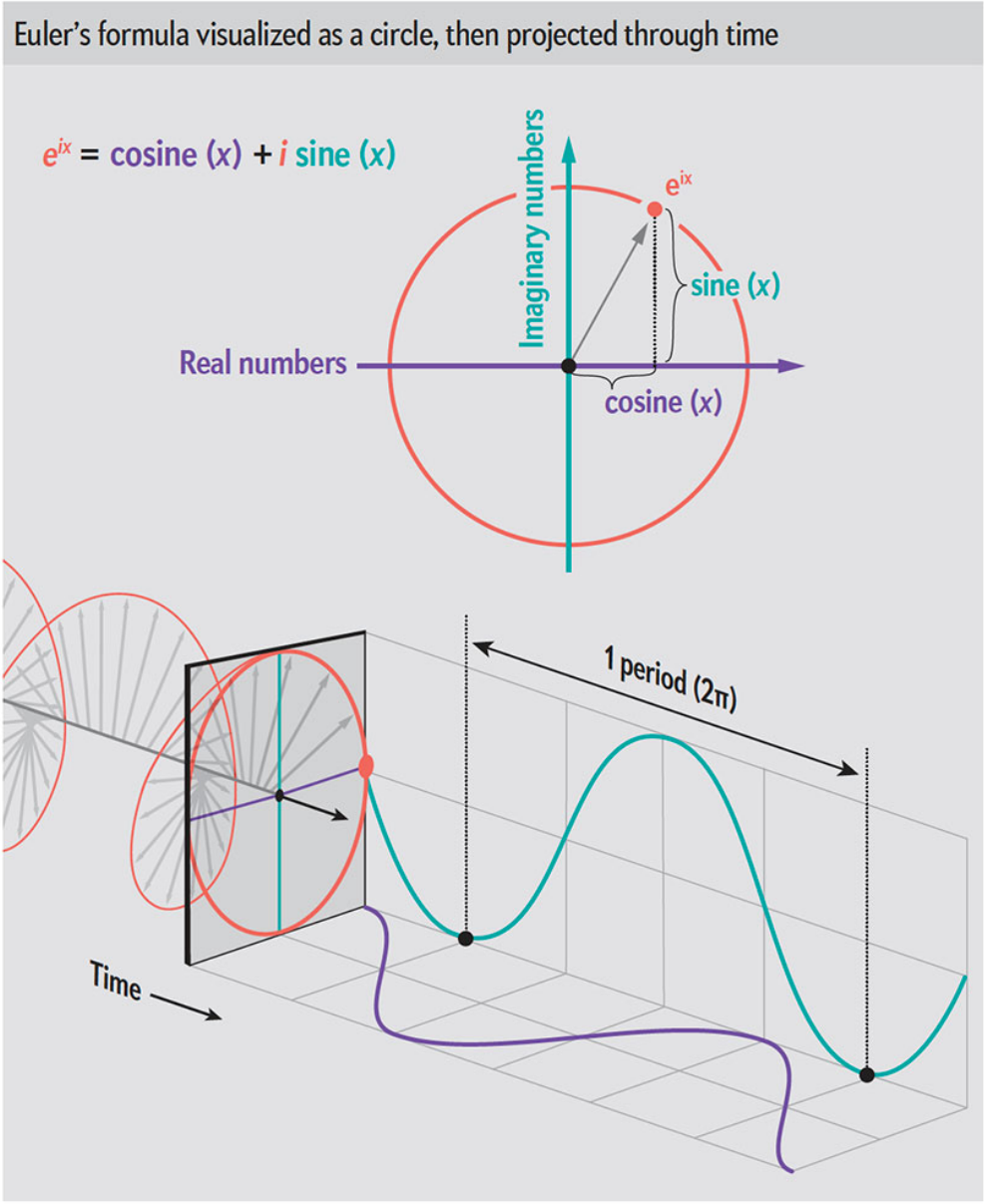
Show a mathematician like Goncharov one of the integrals we get out of Feynman diagrams, and the first thing you will hear is, “That’s a period.”

Periods are a type of number. You might be familiar with the natural numbers (1, 2, 3, 4 . . .) and the rational numbers (fractions). The square root of 2 is not rational—you cannot get it by dividing two natural numbers. What it is, though, is algebraic: you can write an algebraic equation, say $x^2 = 2$, where the square root of 2 is the solution. Periods are the next step up: although you cannot always get them from an algebraic equation, you *can* always get them from an integral.

Why call them periods? In the simplest cases, that is literally what they are: the distance before something repeats. Thinking back to high school, you might remember grappling with sines and cosines. You might even remember that you can put them together with imaginary numbers (the square roots of negative numbers—in other words, numbers that would not normally exist) using Euler's formula: $e^{ix} = \cos(x) + i \sin(x)$ (here e is a constant, and i is the square root of -1). All three of these—sin (x), cos (x), and e^{ix}—have *period* 2Π: if you let x go from 0 to 2Π, the function repeats, and you get the same numbers again.

2Π is a period because it is the distance before e^{ix} repeats, but you can also think of it as an integral. Draw a graph of e^{ix} in the complex plane: imaginary numbers on one axis; real numbers on the other. It forms a circle. If you want to measure the length of that circle, you can do it with an integral, adding up each little segment all the way around. In doing so, you will find exactly 2Π.





Credit: Illustration by Jen Christiansen.

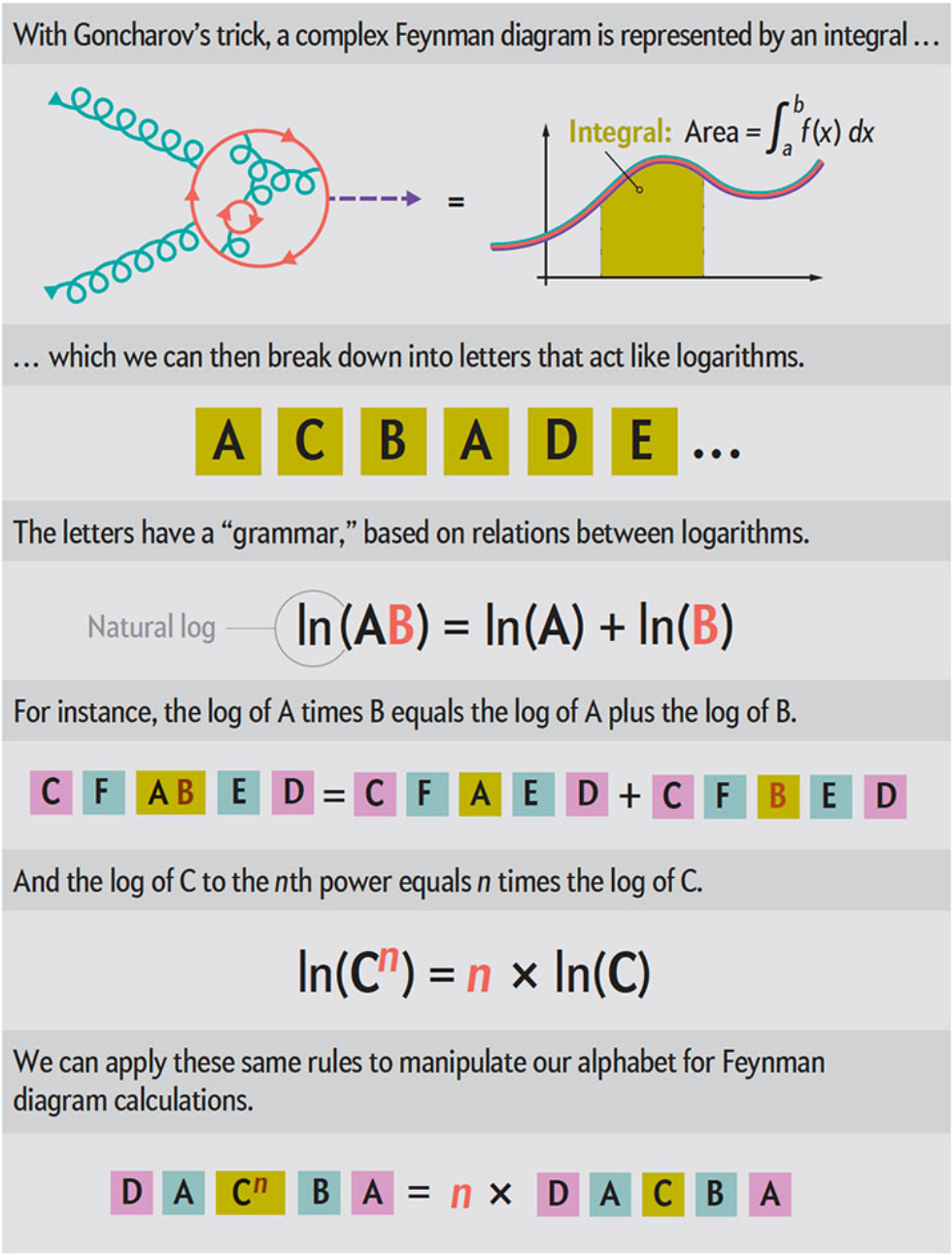
What happens if you go partway around the circle, to some point z? In that case, you must solve the equation $z = e^{ix}$. Thinking back again to high school, you might remember what you need to solve that equation: the natural logarithm, ln (z). Logarithms might not look like "periods" in the way 2Π does, but because you can get them from integrals, mathematicians call them periods as well. Besides 2Π, logarithms are the simplest periods.

The periods mathematicians and physicists care about can be much more complicated than this scenario, of course. In the mid-1990s physicists

started classifying periods in the integrals that come out of Feynman diagrams and have since found a dizzying array of exotic numbers. Remarkably, though, the high school picture remains useful. Many of these exotic numbers, when viewed as periods, can be broken down into logarithms. Understand the logarithms, and you can understand almost everything else.

That was the secret that Goncharov taught Spradlin, Vergu and Volovich. He showed them how to take Del Duca, Duhr and Smirnov's 17-page mess and chop it up into a kind of "alphabet" of logarithms. That alphabet obeys its own "grammar" based on the relations between logarithms, and by using this grammar, the physicists were able to rewrite the result in terms of just a few special "letters," making a messy particle physics calculation look a whole lot simpler.

To recap, physicists calculate scattering amplitudes using Feynman diagrams, which require doing integrals. Those integrals are always periods, sometimes complicated ones, but we can often break those complicated periods apart into simpler periods (logarithms) using Goncharov's trick, which was what ignited my area of the amplitudes field. We can divide many of the integrals we use into an alphabet of letters that behave like logarithms. And the same rules that apply to logarithms, such as basic laws like $\ln(xy) = \ln(x) + \ln(y)$ and $\ln(x^n) = n \times \ln(x)$, work for the alphabet.



Credit: Illustration by Jen Christiansen.

Word Jumble

Goncharov's alphabet trick would not be nearly as impressive if all it did was save space in a journal. Once we know the right alphabet, we can also do new calculations, ones that would not have been possible otherwise. In effect, knowing the alphabet lets us skip the Feynman diagrams and just guess the answer.

Think about that newspaper mainstay, the word jumble. The puzzle tells you which letters you need and how long the word is supposed to be. If you were lazy, you could have a computer write down the letters in every possible order, then skim through the list. Eventually you would find a word that made sense, and you would have your solution.

The list of possibilities can be quite long, though. Luckily in physics, we start with hints. We begin with an alphabet of logarithms that describe the properties our particles can have, such as their energy and speed. Then we start writing words in this alphabet, representing integrals that might show up in the final answer. Certain words do not make physical sense: they describe particles that do not actually exist or diagrams that would be impossible to draw. Others are needed to explain things we already know: what happens when a particle gets very slow or very fast. In the end, we can pare things down from what might have been millions of words to thousands, then tens, and finally just one unique answer. Starting with a guess, we end up with the only possible word that can make sense as our scattering amplitude.

Lance J. Dixon, James M. Drummond and Johannes Henn used this technique to find the right “word” for a three-loop calculation in 2011. I joined the team in 2013, when I snuck away from graduate school on Long Island to spend the winter working for Dixon at SLAC National Accelerator Laboratory at Stanford University. Along with then grad student Jeffrey Pennington, we got the result into a form we could compare with the old two-loop calculation from Del Duca, Duhr and Smirnov. Now instead of 17 pages, we had a formula that was 800 pages long—and all without drawing a single Feynman diagram.

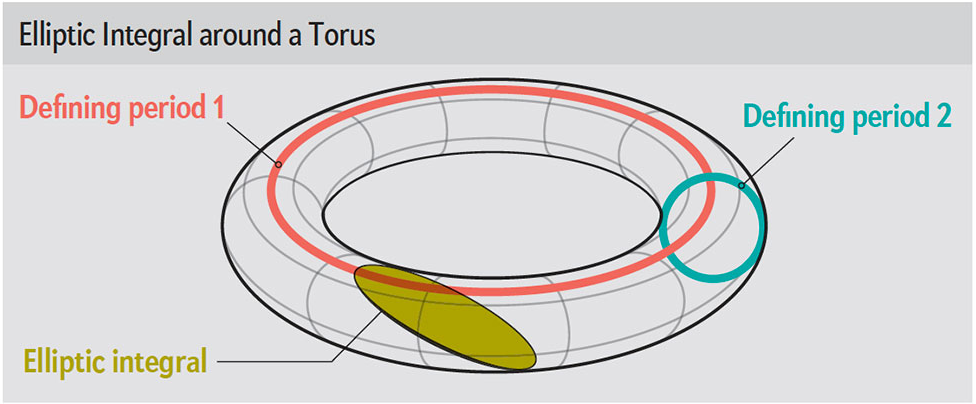
Since then, we have pushed to even more loops, and our collaboration has grown, with Duhr, Andrew McLeod, Simon Caron-Huot, Georgios Papathanasiou and Falko Dulat joining the team. We are at seven loops, and I do not know how many pages the new formulas will take to write out. Goncharov’s trick is not enough to simplify the result when the calculation is this complicated. Here we are just happy it makes the calculation possible! We store our results in computer files now, big enough that you would think they were video files, not text.

The Elliptic Frontier

Recall that the more loops you include in your scattering amplitude calculation, the more precise your prediction will be. Seven loops would be more precise than the two or so loops the LHC can measure, more precise than the four-loop state of the art in quantum electromagnetism. I say "would be" here, though, because there is a catch: our seven-loop calculations use a "toy model"—a simpler theory of particle interactions than any that can describe the real world. Upgrading our calculations so they describe reality will be difficult, and there are numerous challenges. For one, we will need to understand something called elliptic integrals.

The toy model we use is very well behaved. One of its nicer traits is that for the kind of calculations we do, Goncharov's method always works: we can always break the integral up into an alphabet of logarithms, of integrals over circles. In the real world, this tactic runs into problems at two loops: two integrals can get tangled together so they cannot be separated.

Think about two hooked rings that cannot be pulled apart. If you move one ring around the other, you will draw a doughnut shape, or a torus. A torus has two "periods," two different ways you can draw a line around it, corresponding to the two different rings. Integrate around a circle by itself, and you get a logarithm. Try to draw a ring around a torus, and you will not always get a circle: instead you might get an ellipse. We call such integrals around a torus elliptic integrals—integrals over an elliptic curve.



Credit: Illustration by Jen Christiansen.

Understanding elliptic curves involves some famously complex mathematical problems. Some of these problems are so difficult to solve that organizations such as the National Security Agency use them to encode

classified information, on the assumption that no one can solve them fast enough to crack the code. The problems we are interested in are not quite so intractable, but they are still tricky. With the LHC's precision increasing, though, elliptic integrals are becoming more and more essential, spurring on groups around the world to tackle the new mathematics. The machine shut down in late 2018 for upgrades, but scientists still have hordes of data to sort through; it will start up again in 2021 and will go on to produce 10 times more collisions than before.

At times the speed at which the field is moving leaves me breathless. Two winters ago I holed up at Princeton University with a group of collaborators: McLeod, Spradlin, Jacob Bourjaily and Matthias Wilhelm. Within two weeks we went from a sketched-out outline to a full paper, calculating a scattering amplitude involving elliptic integrals. It was the fastest I have ever written a paper, and the entire time we worried that we were going to be scooped, that another group would do the calculation first.

We did not end up getting scooped. But not long after, we received a bit of an early Christmas present: two papers by Duhr, Dulat, Johannes Broedel and Lorenzo Tancredi that explained a better way to handle these integrals, building on work by mathematicians Francis Brown and Andrey Levin. Those papers, along with a later one with Brenda Penante, provided us with the missing piece we needed: a new alphabet of "elliptic letters."

With an alphabet like that, we can apply Goncharov's trick to more complicated integrals and start to understand two-loop amplitudes, not just in a toy model but in the real world as well.

If we can do two-loop calculations in the real world, if we can figure out what the Standard Model predicts to a new level of precision, we will get to see if the LHC's data match those predictions. If it does not, we will have a hint that something genuinely new is going on, something our theories cannot explain. It could be the one piece of data we need to move particle physics to the next frontier, to unlock those lasting mysteries we cannot seem to crack.

--Originally published: Scientific American 320(1); 30-35 (January 2015).

SECTION 4

Mathematics and Human Culture

Geometry v. Gerrymandering

by Moon Duchin

Gerrymandering is clawing across courtrooms and headlines nationwide. The U.S. Supreme Court recently heard cases on the constitutionality of voting districts that allegedly entrenched a strong advantage for Republicans in Wisconsin and Democrats in Maryland but dodged direct rulings in both. Another partisan gerrymandering case from North Carolina is winding its way up with a boost from an emphatic lower court opinion in August of 2018. But so far it has been impossible to satisfy the justices with a legal framework for partisan gerrymandering. Part of the problem, as former justice Anthony Kennedy noted in a 2004 case, is that courts high and low have yet to settle on a "workable standard" for identifying a partisan gerrymander in the first place. That is where a growing number of mathematicians around the country think we can help.

In 2016, with a few friends, I founded a working group to study the applications of geometry and computing to redistricting in the U.S. Since then, the Metric Geometry and Gerrymandering Group has expanded its scope and mission, becoming deeply engaged in research, outreach, training and consulting. More than 1,200 people have attended our workshops around the country, and many of them have become intensely involved in redistricting projects. We think the time is right to make a computational intervention. The mathematics of gerrymandering is surprisingly rich—enough to launch its own subfield—and computing power is arguably just catching up with the scale and complexity of the redistricting problem. Despite our group's technical orientation, our central goal is to reinforce and protect civil rights, and we are working closely with lawyers, political scientists, geographers and community groups to build tools and ideas in advance of the next U.S. Census and the round of redistricting to follow it.

In a country that vests power in elected representatives, there will always be skirmishes for control of the electoral process. And in a system such as that of our House of Representatives—where winner takes all within each geographical district—the delineation of voting districts is a natural battleground. American history is chock-full of egregious line-drawing schemes, from stung a district with an incumbent's loyalists to slicing a long-standing district three ways to suppress the political power of black voters. Many varieties of these so-called packing and cracking strategies continue today, and in the big data moment, they have grown enormously more sophisticated. Now more than ever, abusive redistricting is stubbornly difficult to even identify definitively. People think they know gerrymandering by two hallmarks—bizarre shapes and disproportionate electoral outcomes—yet neither one is reliable. So how do we determine when the scales are unfairly tipped?

The Eyeball Test

The 1812 episode that gave us the word "gerrymander" sprang from the intuition that oddly shaped districts betray an illegitimate agenda. It is named for Elbridge Gerry, who was governor of Massachusetts at the time. Gerry had quite a Founding Father pedigree—signer of the Declaration of Independence, major player at the U.S. Constitutional Convention, member of Congress, James Madison's vice president—so it is amusing to consider that his enduring fame comes from nefarious redistricting. "Gerry-mander," or Gerry's salamander, was the satirical name given to a curvy district in Boston's North Shore that was thought to favor the governor's Democratic-Republican party over the rival Federalists. A woodcut political cartoon ran in the *Salem Gazette* in 1813; in it, wings, claws and fangs were suggestively added to the district's contours to heighten its appearance of reptilian contortions.

So the idea that erratic districts tip us off to wrongdoing goes a long way back, and the converse notion that close-knit districts promote democratic ideals is as old as the republic. In 1787 Madison wrote in *The Federalist Papers* that "the natural limit of a democracy is that distance from the central point which will just permit the most remote citizens to assemble as often as their public functions demand. " In other words, districts should be transitable. In 1901 a federal apportionment act marked the first appearance

in U.S. law of the vague desideratum that districts should be composed of “compact territory. ” The word “compact” then proliferated throughout the legal landscape of redistricting but almost always without a definition.

For instance, at a 2017 meeting of the National Conference of State Legislatures, I learned that after the last Census, Utah’s lawmakers took the commendable time and effort to set up a Web site, Redistrict Utah, to solicit proposed districting maps from everyday citizens. To be considered, maps were required to be “reasonably compact.” I jumped at the opportunity to find out how exactly that quality was being tested and enforced, only to learn that it was handled by just tossing the funny-looking maps. If that sounds bad, Utah is far from alone. Thirty-seven states have some kind of shape regulation on the books, and in almost every case, the eyeball test is king.

The problem is that the outline of a district tells a very partial and often misleading story. On one hand there can certainly be benign reasons for ugly shapes. Physical geography or reasonable attempts to follow county lines or unite communities of interest can influence a boundary, although just as often, legitimate priorities such as these are merely scapegoated in an attempt to defend the worst-off ending districts. On the other hand districts that are plump, squat and symmetrical offer no meaningful seal of quality. In 2018 a congressional redistricting plan in Pennsylvania drafted by Republicans in the state legislature achieved strong compactness scores under all five formulas specified by Pennsylvania’s supreme court. Yet mathematical analysis revealed that the plan would nonetheless lock in the same extreme partisan skew as the contorted plan, enacted in 2011, that it was meant to replace. So the justices opted for the extraordinary measure of adopting an independent outsider’s plan.

Lopsided Outcomes

If shape is not a reliable indicator of gerrymandering, what about studying the extent to which elected representatives match the voting patterns of the electorate? Surely lopsided outcomes provide prima facie evidence of abuse. But not so fast. Take Republicans in my home state of Massachusetts. In the 13 federal elections for president and Senate since 2000, GOP candidates have averaged more than one third of the votes statewide. That is six times the level needed to win a seat in one of

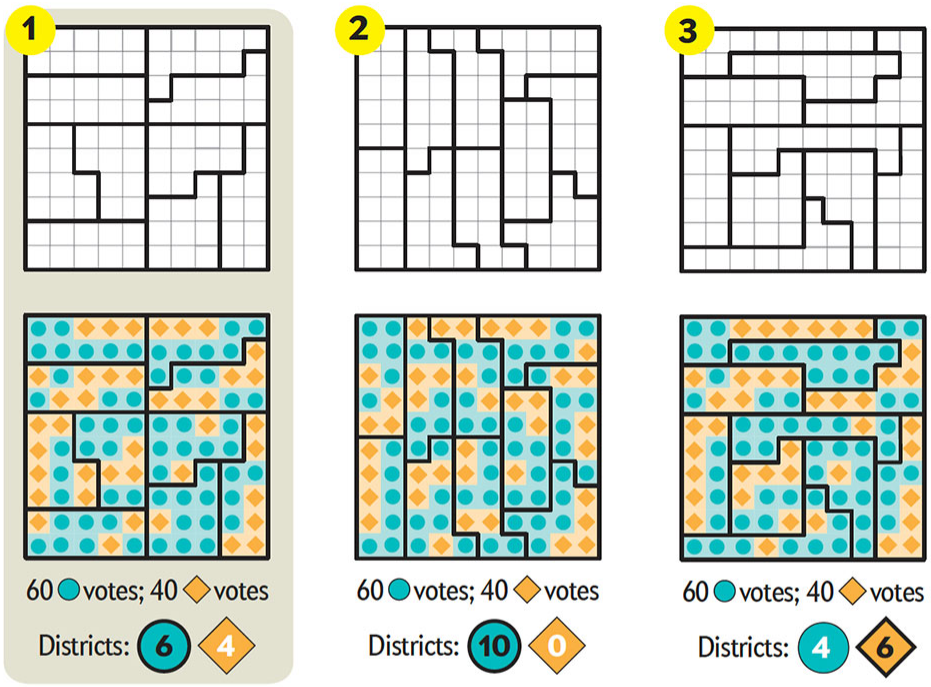
Massachusetts's nine congressional districts because a candidate in a two-way race needs a simple majority to win. Yet no Republican has won a seat in the House since 1994.

We must be looking at a gerrymander that denies Republicans their rightful opportunity districts, right? Except the mathematics here is completely exonerating. Let us look at a statewide race so that we can put uncontested seats and other confounding variables to the side. Take Kenneth Chase, the Republican challenger to Ted Kennedy for the U.S. Senate in 2006, who cracked 30 percent of the statewide vote. Proportionally, you would expect Chase to beat Kennedy in nearly three out of nine congressional districts. But the numbers do not shake out. As it turns out, it is mathematically impossible to select a single district-sized grouping of towns or precincts, even scattered around the state, that preferred Chase. His voters simply were not clustered enough. Instead most precincts went for Chase at levels close to the state average, so there were too few Chase-favoring building blocks to go around.

Any voting minority needs a certain level of nonuniformity in how its votes are distributed for our districting system to offer even a theoretical opportunity to secure representation. And the type of analysis applied to the Chase-Kennedy race does not even consider spatial factors, such as the standard requirement that each district be one connected piece. One may rightfully wonder how we can ever hold district architects accountable when the landscape of possibilities can hold so many surprises.

The Power of the Pen

Gerrymandering relies on carefully drawn lines that dilute the voting power of one population to favor another by clustering one side's voters into a few districts with excessively high numbers (packing), by dispersing them across several districts so that they fall short of electing a preferred candidate (cracking), or by using a combination of the two schemes.



A grid is districted to produce an electoral outcome proportional to the share of votes for each party 1. The same grid can be districted using combinations of packing and cracking to produce extreme outcomes 2, 3—one in which the Blue party wins all districts and one in which it wins only four of 10. In this particular case, the geometry of the layout turns out to favor the Blue party. Statistical analysis using Markov chain Monte Carlo reveals that the Orange party is far more likely to get two or three seats, rather than its proportional share of four, in the universe of possible plans.

Credit: Illustration by Jen Christiansen.

Random Walks to the Rescue

The only reasonable way to assess the fairness of a districting plan is to compare it with other valid plans for cutting up the same jurisdiction because you must control for aspects of electoral outcomes that were forced by the state's laws, demographics and geography. The catch is that studying the universe of possible plans becomes an intractably big problem.

Think of a simple four-by-four grid and suppose you want to divide it into four contiguous districts of equal size, with four squares each. If we imagine the grid as part of a chessboard, and we interpret contiguity to mean that a rook should be able to visit the entire district, then there are exactly 117 ways to do it. If corner adjacency is permitted—so-called queen contiguity—then there are 2,620 ways. And they are not so straightforward to count. As my colleague Jim Propp, a professor at the University of

Massachusetts Lowell and a leader in the field of combinatorial enumeration, puts it, “In one dimension, you can split paths along the way to divide and conquer, but in two dimensions, suddenly there are many, many ways to get from point A to point B.”

The issue is that the best counting techniques often rely on recursion—that is, solving a problem using a similar problem that is a step smaller—but two-dimensional spatial counting problems just do not recurse well without some extra structure. So complete enumerations must rely on brute force. Whereas a cleverly programmed laptop can classify partitions of small grids nearly instantly, we see huge jumps in complexity as the grid size grows, and the task quickly zooms out of reach. By the time you get to a grid of nine-by-nine, there are more than 700 trillion solutions for equinumerous rook partitions, and even a high-performance computer needs a week to count them all. This seems like a hopeless state of affairs. We are trying to assess one way of cutting up a state without any ability to enumerate—let alone meaningfully compare it against—the universe of alternatives. This situation sounds like groping around in a dark, infinite wilderness.

The good news is that there is an industry standard used across scientific domains for just such a colossal task: Markov chain Monte Carlo (MCMC). Markov chains are random walks in which where you go next is governed by probability, depending only on where you are now (at every position, you roll the dice to choose a neighboring space to move to). Monte Carlo methods are just estimation by random sampling. Put them together, and you get a powerful tool for searching vast spaces of possibilities. MCMC has been successfully used to de code prison messages, probe the properties and phase transitions of liquids, find provably accurate fast approximations for hard computational problems, and much more. A 2009 survey by the eminent statistician Persi Diaconis estimated that MCMC drives 10 to 15 percent of the statistical work in science, engineering and business, and the number has probably only gone up since then. Although computational analysis in redistricting goes back several decades, serious attempts to apply MCMC in that effort only started to appear publicly around 2014.

Imagine that officials in the state of Gridlandia hire you to decide if their legislature’s districting plan is reasonable. If Gridlandia is a four-by-four

grid of squares, and its state constitution calls for rook-contiguous districts, then you are in luck: there are exactly 117 ways to produce a compliant plan, and you can examine them all. You can set up a perfectly faithful model of this universe of districting plans by using 117 nodes to represent the valid plans and adding edges between the nodes to represent simple moves in which two squares in the grid swap their district assignments. The edges give you a way of conceptualizing how similar two plans are by simply counting the number of swaps needed to transform one to the other. (I call this structure a “metagraph” because it is a graph of ways to cut up another graph.) Now suppose that the state legislature is controlled by the Diamond party, and its rivals suspect that it has rigged the seats in its favor. To determine if that is true, one may turn to the election data. If the Diamond plan would have produced more seats for the party in the last election than, say, 114 out of 117 alternatives and if the same is true for several previous elections, the plan is clearly a statistical outlier. This is persuasive evidence of a partisan gerrymander—and you do not need MCMC for such an analysis.

The MCMC method kicks in when you have a full-sized problem in place of this small toy problem. As soon as you get past 100 or so nodes, there is a similar metagraph, but you cannot completely build it because of its forbidding complexity. That is no deal breaker, though. From any single plan, it is still easy to build out the local neighborhood by performing all possible moves. Now you can take a million, billion or trillion steps and see what you find. There is mathematics in the background (ergodic theory, to be precise) guaranteeing that if you random-walk for long enough, the ensemble of maps you collect will have properties representative of the overall universe, typically long before you have visited even a modest fraction of nodes in your state space. This lets you determine if the map you are evaluating is an extreme outlier according to various partisan metrics.

The cutting edge of scientific inquiry is to build more powerful algorithms and, at the same time, to devise new theorems that certify that we are sampling well enough to draw robust conclusions. There is an emerging scientific consensus around this method but also many directions of ongoing research.

Markov chains are random walks around a graph or network in which the next destination is determined by a probability, like a roll of the dice, depending on the current position. Monte Carlo methods use random sampling to estimate a distribution of probabilities. Combined, Markov chain Monte Carlo (MCMC) is a powerful tool for searching and sampling from a vast space of scenarios, such as all the possible districting plans in a state. Attempts to use computational analysis to spot devious districting go back several decades, but efforts to apply MCMC to the problem are much more recent.

Dimensions; Districts	Equal-size districts	District sizes can be unequal (+/- 1)
2×2 grid; 2 districts	2	6
3×3 grid; 3 districts	10	58
4×4 grid; 2 districts	70	206
4×4 grid; 4 districts	117	1,953
4×4 grid; 8 districts	36	34,524
5×5 grid; 5 districts	4,006	193,152
6×6 grid; 2 districts	80,518	?*
6×6 grid; 3 districts	264,500	?
6×6 grid; 4 districts	442,791	?
6×6 grid; 6 districts	451,206	?
6×6 grid; 9 districts	128,939	?
6×6 grid; 12 districts	80,092	?
6×6 grid; 18 districts	6,728	?
7×7 grid; 7 districts	158,753,814	?
8×8 grid; 8 districts	187,497,290,034	?
9×9 grid; 9 districts	706,152,947,468,301	?

SIMPLE CASE

It is easy to enumerate all the ways to partition a small grid into equal-size districts. For a two-by-two grid with two districts of equal size, there are only two solutions. But if districts can vary in size, the number of solutions jumps to six.

Equal-size districts: 2 solutions

District size can be +/-1: 6 solutions

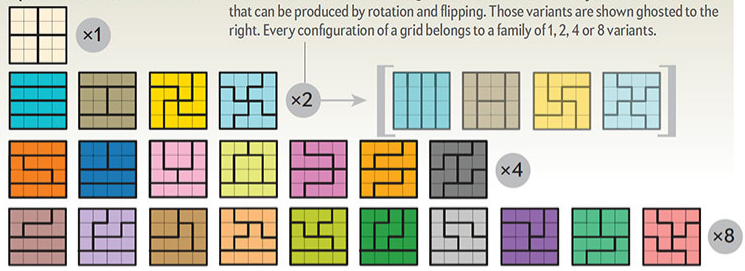
Mathematicians have not yet enumerated these solutions, which can require a week of computing or more. To find out more about the hunt for these numbers, visit www.mggg.org

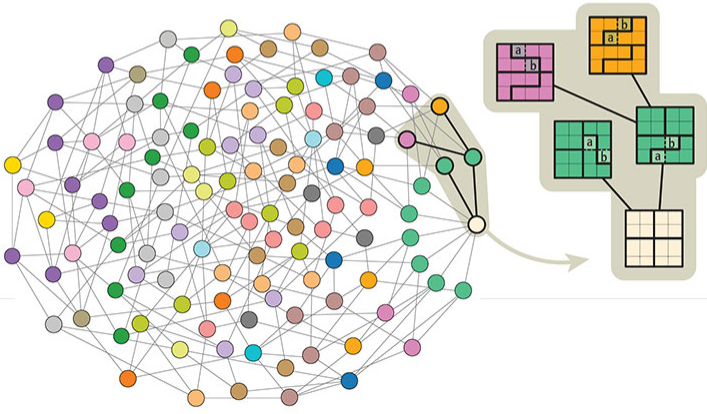
BIGGER CASE

As the size of the grid grows, the number of possibilities for carving it up skyrockets. Dividing a four-by-four grid into four districts of equal size has 117 solutions. If the districts can vary in size by even one unit, there are 1,953 solutions. It does not take long before even the most powerful computers struggle to enumerate the possibilities for more complex grids. That presents a problem for anyone trying to detect manipulative maps by comparing the myriad ways to district a U.S. state. But MCMC can help.

Equal-size districts: 117 solutions

"×2" means that each of the configurations on the left has exactly one distinct variation that can be produced by rotation and flipping. Those variants are shown ghosted to the right. Every configuration of a grid belongs to a family of 1, 2, 4 or 8 variants.





We can efficiently explore valid districting plans by traveling randomly around a "metagraph," defined by moves such as the unit swaps pictured. In the highlighted inset, each pattern has squares marked a and b whose district assignments are exchanged to arrive at the configuration of the pattern shown. The edges in the network represent these simple swap moves. The metagraph models the space of all valid districting plans and can be used to sample many billions of plans. Geometers are trying to understand the shape and structure of that universe of plans.

Credit: Illustration by Jen Christiansen.

So far courts seem to be smiling on this approach. Two mathematicians—Duke University's Jonathan Mattingly and Carnegie Mellon University's Wes Pegden—have recently testified about MCMC approaches for the federal case in North Carolina and the state-level case in Pennsylvania, respectively.

Mattingly used MCMC to characterize the reasonable range one might observe for various metrics, such as seats won, across ensembles of districting plans. His random walk was weighted to favor plans that were deemed closer to ideal, along the lines of North Carolina state law. Using his ensembles, he argued that the enacted plan was an extreme partisan outlier. Pegden used a different kind of test, appealing to a rigorous theorem that quantifies how unlikely it is that a neutral plan would score much worse than other plans visited by a random walk. His method produces p-values, which constrain how improbable it is to find such anomalous bias by chance. Judges found both arguments credible and cited them favorably in their respective decisions.

For my part, Pennsylvania governor Tom Wolf brought me on earlier this year as a consulting expert for the state's scramble to draw new district lines following its supreme court's decision to strike down the 2011 Republican plan. My contribution was to use the MCMC framework to evaluate new plans as they were proposed, harnessing the power of statistical outliers while adding new ways to take into account more of the varied districting principles in play, from compactness to county splits to community structure. My analysis agreed with Pegden's in flagging the 2011 plan as an extreme partisan outlier—and I found the new plan floated by the legislature to be just as extreme, in a way that was not explained away by its improved appearances.

As the 2020 Census approaches, the nation is bracing for another wild round of redistricting, with the promise of litigation to follow. I hope the next steps will play out not just in the courtrooms but also in reform measures that require a big ensemble of maps made with open-source tools to be examined before any plan gets signed into law. In that way, the legislatures preserve their traditional prerogatives to commission and

approve district boundaries, but they have to produce some guarantees that they are not putting too meaty a thumb on the scale.

Computing will never make tough redistricting decisions for us and cannot produce an optimally fair plan. But it can certify that a plan behaves as though selected just from the stated rules. That alone can rein in the worst abuses and start to restore trust in the system.

--Originally published: Scientific American 319(5); 48-53 (November 2018).

Hacking Passwords with Math

by Jean-Paul Delahaye

At one time or another, we have all been frustrated by trying to set a password, only to have it rejected as too weak. We are also told to change our choices regularly. Obviously such measures add safety, but how exactly?

I will explain the mathematical rationale for some standard advice, including clarifying why six characters are not enough for a good password and why you should never use only lowercase letters. I will also explain how hackers can uncover passwords even when stolen data sets lack them.

ChOose#W!sely@*

Here is the logic behind setting hack-resistant passwords. When you are asked to create a password of a certain length and combination of elements, your choice will fit into the realm of all unique options that conform to that rule—into the "space" of possibilities. For example, if you were told to use six lowercase letters—such as, afzjxd, auntie, secret, wwwwww—the space would contain 26^6, or 308,915,776, possibilities. In other words, there are 26 possible choices for the first letter, 26 possible choices for the second, and so forth. These choices are independent: you do not have to use different letters, so the size of the password space is the product of the possibilities, or 26 x 26 x 26 x 26 x 26 x 26 = 26^6.

If you are told to select a 12-character password that can include uppercase and lowercase letters, the 10 digits and 10 symbols (say, !, @, #, $, %, ^, &, ?, / and +), you would have 72 possibilities for each of the 12 characters of the password. The size of the possibility space would then be 72^{12} (19,408,409,961,765,342,806,016, or close to 19 x 10^{21}).

That is more than 62 trillion times the size of the first space. A computer running through all the possibilities for your 12-character password one by one would take 62 trillion times longer. If your computer spent a second visiting the six-character space, it would have to devote two million years to examining each of the passwords in the 12-character space. The multitude of possibilities makes it impractical for a hacker to carry out a plan of attack that might have been feasible for the six-character space.

Calculating the size of these spaces by computer usually involves counting the number of binary digits in the number of possibilities. That number, *N*, is derived from this formula: 1 + integer($\log_2(N)$). In the formula, the value of $\log_2(N)$ is a real number with many decimal places, such as $\log_2(26^6) = 28.202638\ldots$. The "integer" in the formula indicates that the decimal portion of that log value is omitted, rounding down to a whole number—as in integer(28.202638… 28). For the example of six lowercase letters above, the computation results in 29 bits; for the more complex, 12-character example, it is 75 bits. (Mathematicians refer to the possibility spaces as having entropy of 29 and 75 bits, respectively.) The French National Cybersecurity Agency (ANSSI) recommends spaces having a minimum of 100 bits when it comes to passwords or secret keys for encryption systems that absolutely must be secure. Encryption involves representing data in a way that ensures it cannot be retrieved unless a recipient has a secret code-breaking key. In fact, the agency recommends a possibility space of 128 bits to guarantee security for several years. It considers 64 bits to be very small (very weak); 64 to 80 bits to be small; and 80 to 100 bits to be medium (moderately strong).

Moore's law (which says that the computer-processing power available at a certain price doubles roughly every two years) explains why a relatively weak password will not suffice for long-term use: over time computers using brute force can find passwords faster. Although the pace of Moore's law appears to be decreasing, it is wise to take it into account for passwords that you hope will remain secure for a long time.

For a truly strong password as defined by ANSSI, you would need, say, a sequence of 16 characters, each taken from a set of 200 characters. This would make a 123-bit space, which would render the password close to impossible to memorize. Therefore, system designers are generally less

demanding and accept low- or medium-strength passwords. They insist on long ones only when the passwords are automatically generated by the system, and users do not have to remember them.

There are other ways to guard against password cracking. The simplest is well known and used by credit cards: after three unsuccessful attempts, access is blocked. Alternative ideas have also been suggested, such as doubling the waiting time after each successive failed attempt but allowing the system to reset after a long period, such as 24 hours. These methods, however, are ineffective when an attacker is able to access the system without being detected or if the system cannot be configured to interrupt and disable failed attempts.

How Long Does It Take to Search All Possible Passwords?

For a password to be difficult to crack, it should be chosen randomly from a large set, or "space," of possibilities. The size, $\boldsymbol{T}$, of the possibility space is based on the length, $\boldsymbol{A}$, of the list of valid characters in the password and the number of characters, $\boldsymbol{N}$, in the password. The size of this space ($\boldsymbol{T} = \boldsymbol{A}^{\boldsymbol{N}}$) may vary considerably.

Each of the following examples specifies values of $\boldsymbol{A}$, $\boldsymbol{N}$, $\boldsymbol{T}$ and the number of hours, $\boldsymbol{D}$, that hackers would have to spend to try every permutation of characters one by one. $\boldsymbol{X}$ is the number of years that will have to pass before the space can be checked in less than one hour, assuming that Moore's law (the doubling of computing capacity every two years) remains valid. I also assume that in 2019, a computer can explore a billion possibilities per second. I represent this set of assumptions with the following three relationships and consider five possibilities based on values of $\boldsymbol{A}$ and $\boldsymbol{N}$:

Relationships

$$\boldsymbol{T} = \boldsymbol{A}^{\boldsymbol{N}}$$
$$\boldsymbol{D} = \boldsymbol{T}/(10^9 \times 3{,}600)$$
$$\boldsymbol{X} = 2 \log_2[\boldsymbol{T}/(10^9 \times 3{,}600)]$$

Results

If $A = 26$ and $N = 6$, then $T = 308{,}915{,}776$
$D = 0.0000858$ computing hour
$X = 0$; it is already possible to crack all passwords in the space in under an hour

If $A = 26$ and $N = 12$, then $T = 9.5 \times 10^{16}$
$D = 26{,}508$ computing hours
$X = 29$ years before passwords can be cracked in under an hour

If $A = 100$ and $N = 10$, then $T = 10^{20}$
$D = 27{,}777{,}777$ computing hours
$X = 49$ years before passwords can be cracked in under an hour

If $A = 100$ and $N = 15$, then $T = 10^{30}$
$D = 2.7 \times 10^{17}$ computing hours
$X = 115$ years before passwords can be cracked in under an hour

If $A = 200$ and $N = 20$, then $T = 1.05 \times 10^{46}$
$D = 2.7 \times 10^{33}$ computing hours
$X = 222$ years before passwords can be cracked in under an hour

Weaponizing Dictionaries and Other Hacker Tricks

Quite often an attacker succeeds in obtaining encrypted passwords or password "fingerprints" (which I will discuss more fully later) from a system. If the hack has not been detected, the interloper may have days or even weeks to attempt to derive the actual passwords.

To understand the subtle processes exploited in such cases, take another look at the possibility space. When I spoke earlier of bit size and password space (or entropy), I implicitly assumed that the user consistently chooses passwords at random. But typically the choice is not random: people tend to select a password they can remember (locomotive) rather than an arbitrary string of characters (xdichqewax).

This practice poses a serious problem for security because it makes passwords vulnerable to so-called dictionary attacks. Lists of commonly used passwords have been collected and classified according to how frequently they are used. Attackers attempt to crack passwords by going through these lists systematically. This method works remarkably well

because, in the absence of specific constraints, people naturally choose simple words, surnames, first names and short sentences, which considerably limits the possibilities. In other words, the nonrandom selection of passwords essentially reduces possibility space, which decreases the average number of attempts needed to uncover a password.

Below are the first 25 entries in one of these password dictionaries, listed in order, starting with the most common one. (I took the examples from a database of five million passwords that was leaked in 2017 and analyzed by SplashData.)

1. 123456
2. password
3. 12345678
4. qwerty
5. 12345
6. 123456789
7. letmein
8. 1234567
9. football
10. iloveyou
11. admin
12. welcome
13. monkey
14. login
15. abc123
16. starwars
17. 123123
18. dragon
19. passw0rd
20. master
21. hello
22. freedom
23. whatever
24. qazwsx
25. trustno1

If you use password or iloveyou, you are not as clever as you thought! Of course, lists differ according to the country where they are collected and the Web sites involved; they also vary over time.

For four-digit passwords (for example, the PIN code of SIM cards on smartphones), the results are even less imaginative. In 2013, based on a collection of 3.4 million passwords each containing four digits, the DataGenetics Web site reported that the most commonly used four-digit sequence (representing 11 percent of choices) was 1234, followed by 1111 (6 percent) and 0000 (2 percent). The least-used four-digit password was 8068. Careful, though, this ranking may no longer be true now that the result has been published. The 8068 choice appeared only 25 times among the 3.4-million four-digit sequences in the database, which is much less than the 340 uses that would have occurred if each four-digit combination had been used with the same frequency. The first 20 series of four digits are: 1234; 1111; 0000; 1212; 7777; 1004; 2000; 4444; 2222; 6969; 9999; 3333; 5555; 6666; 1122; 1313; 8888; 4321; 2001; 1010.

Even without a password dictionary, using differences in frequency of letter use (or double letters) in a language makes it possible to plan an effective attack. Some attack methods also take into account that, to facilitate memorization, people may choose passwords that have a certain structure—such as A1=B2=C3, AwX2AwX2 or O0o.lli. (which I used for a long time)—or that are derived by combining several simple strings, such as password123 or johnABC0000. Exploiting such regularities makes it possible to for hackers to speed up detection.

Making Hash of Hackers

As the main text explains, instead of storing clients' passwords, Internet servers store the "fingerprints" of these passwords: sequences of characters that are derived from the passwords. In the event of an attack, the use of fingerprints can make it is very difficult, if not impossible, for hackers to use what they find.

The transformation is achieved by using algorithms known as cryptographic hash functions. These are meticulously developed processes that transform a data file, F, however long it may be, into a sequence, h(F),

called a fingerprint of F. For example, the hash function SHA256 transforms the phrase “Nice weather” into:

DB0436DB78280F3B45C2E09654522197D59EC98E7E64AEB967A2A 19EF7C394A3 (64 hexadecimal, or base 16, characters, which is equivalent to 256 bits)

Changing a single character in the file completely alters its fingerprint. For example, if the first character of Nice weather is changed to lowercase (nice weather), the hash SHA256 will generate another fingerprint:

02C532E7418CD1B57961A1B090DB6EC37B3C58380AC0E6877F3B6 155C974647E

You can do these calculations yourself and check them at https://passwordsgenerator.net/sha256-hash-generator or www.xorbin.com/tools/sha256-hash-calculator

Good hash functions produce fingerprints that are similar to those that would be obtained if the fingerprint sequence was uniformly chosen at random. In particular, for any possible random result (a sequence of 64 hexadecimal characters), it is impossible to find a data file F with this fingerprint in a reasonable amount of time.

There have been several generations of hash functions. The SHA0 and SHA1 generations are obsolete and are not recommended. The SHA2 generation, including SHA256, is considered secure.

The Take-Home For Consumers

Taking all this into account, properly designed Web sites analyze the passwords proposed at the time of their creation and reject those that would be too easy to recover. It is irritating, but it’s for your own good.

The obvious conclusion for users is that they must choose their passwords randomly. Some software does provide a random password. Be aware, however, that such password-generating software may, deliberately or not, use a poor pseudo-random generator, in which case what it provides may be imperfect.

You can check whether any of your passwords has already been hacked by using a Web tool called Pwned Passwords

(https://haveibeenpwned.com/Passwords). Its database includes more than 500 million passwords obtained after various attacks.

I tried e=mc2e=mc2, which I liked and believed to be secure, and received an unsettling response: “This password has been seen 114 times before.” Additional attempts show that it is difficult to come up with easy-to-memorize passwords that the database does not know. For example, aaaaaa appeared 395,299 times; a1b2c3d4, 113,550 times; abcdcba, 378 times; abczyx, 186 times; acegi, 117 times; clinton, 18,869 times; bush, 3,291 times; obama, 2,391 times; trump, 859 times.

It is still possible to be original. The Web site did not recognize the following six passwords, for example: eyahaled (my name spelled backward); bizzzzard; meaudepace and modeuxpass (two puns on the French for “password”); abcdef2019; passwaurde. Now that I’ve tried them, I wonder if the database will add them when it next updates. In that case, I won’t use them.

Advice For Web Sites

Web sites, too, follow various rules of thumb. The National Institute of Standards and Technology recently published a notice recommending the use of dictionaries to filter users’ password choices.

Among the rules that a good Web server designer absolutely must adhere to is, do not store plaintext lists of usernames and passwords on the computer used to operate the Web site.

The reason is obvious: hackers could access the computer containing this list, either because the site is poorly protected or because the system or processor contains a serious flaw unknown to anyone except the attackers (a so-called zero-day flaw), who can exploit it.

One alternative is to encrypt the passwords on the server: use a secret code that transforms them via an encryption key into what will appear to be random character sequences to anyone who does not possess the decryption key. This method works, but it has two disadvantages. First, it requires decrypting the stored password every time to compare it with the user’s entry, which is inconvenient. Second, and more seriously, the decryption necessary for this comparison requires storing the decryption key in the

Web site computer's memory. This key may therefore be detected by an attacker, which brings us back to the original problem.

A better way to store passwords is through what are called hash functions that produce "fingerprints." For any data in a file—symbolized as F—a hash function generates a fingerprint. (The process is also called condensing or hashing.) The fingerprint—*h*(F)—is a fairly short word associated with F but produced in such a way that, in practice, it is impossible to deduce F from *h*(F). Hash functions are said to be one-way: getting from F to *h*(F) is easy; getting from *h*(F) to F is practically impossible. In addition, the hash functions used have the characteristic that even if it is possible for two data inputs, F and F', to have the same fingerprint (known as a collision), in practice for a given F, it is almost impossible to find an F' with a fingerprint identical to F.

Using such hash functions allows passwords to be securely stored on a computer. Instead of storing the list of paired usernames and passwords, the server stores only the list of username/fingerprint pairs.

When a user wishes to connect, the server will read the individual's password, compute the fingerprint and determine whether it corresponds to the list of stored username/fingerprint pairs associated with that username. That maneuver frustrates hackers because even if they have managed to access the list, they will be unable to derive the users' passwords, inasmuch as it is practically impossible to go from fingerprint to password. Nor can they generate another password with an identical fingerprint to fool the server because it is practically impossible to create collisions.

Still, no approach is foolproof, as is highlighted by frequent reports of the hacking of major sites. In 2016, for example, data from a billion accounts were stolen from Yahoo!

For added safety, a method known as salting is sometimes used to further impede hackers from exploiting stolen lists of username/fingerprint pairs. Salting is the addition of a unique random string of characters to each password. It ensures that even if two users employ the same password, the stored fingerprints will differ. The list on the server will contain three components for each user: username, fingerprint derived after salt was added to the password, and the salt itself. When the server checks the

password entered by a user, it adds the salt, computes the fingerprint and compares the result with its database.

Even when user passwords are weak, this method considerably complicates the hacker's work. Without salting, a hacker can compute all the fingerprints in a dictionary and see those in the stolen data; all the passwords in the hacker's dictionary can be identified. With salting, for every salt used, the hacker must compute the salted fingerprints of all the passwords in the hacker's dictionary. For a set of 1,000 users, this multiplies by 1,000 the computations required to use the hacker's dictionary.

Survival of the Fittest

It goes without saying that hackers have their own ways of fighting back. They face a dilemma, though: their simplest options either take a lot of computing power or a lot of memory. Often neither option is viable. There is, however, a compromise approach known as the rainbow table method.

In the age of the Internet, supercomputers and computer networks, the science of password setting and cracking continues to evolve—as does the relentless struggle between those who strive to protect passwords and those who are determined to steal, and potentially abuse, them.

Rainbow Tables Help Hackers

Say you are a hacker looking to exploit data that you have acquired. These data consist of username/fingerprint pairs, and you know the hash function. The password is contained in the possibility space of strings of 12 lowercase letters, which corresponds to 56 bits of information and 26^{12} (9.54×10^{16}) possible passwords.

At least two strong approaches are open to you:

Method 1. You scroll through the entire space of passwords. You calculate the fingerprint, $h(P)$, for each password, checking to see whether it appears in the stolen data. You do not need a lot of memory, because prior results are deleted with each new attempt, although you do, of course, have to keep track of the possibilities that have been tested.

Scrolling through all the possible passwords in this way takes a long time. If your computer runs a billion tests per second, you will need $26^{12}/(10^9 \times 3{,}600 \times 24)$ days (1,104 days), or about three years to complete the task. The feat is not impossible; if you happen to have a computer network of 1,000 machines, one day will suffice. It is not feasible, however, to repeat such a calculation every time you wish to test additional data, such as if you obtain a new set of username/fingerprint pairs. (Because you have not saved the results of your computations, you would need an additional 1,104 days to process the new information.)

Method 2. You say to yourself, "I'll compute the fingerprints of all possible passwords, which will take time, and I'll store the resulting fingerprints in a big table. Then I'll have to find only a password fingerprint in the table to identify the corresponding password in the stolen data."

You will need $(9.54 \times 10^{16}) \times (12 + 32)$ bytes of memory because the task requires 12 bytes for the password and 32 bytes for the fingerprint if the fingerprint contains 256 bits (assuming an SHA256 function). That's 4.2×10^{18} bytes, or 4.2 million hard disks with a capacity of one terabyte.

This memory requirement is too large. Method 2 is no more feasible than method 1. Method 1 requires too many computations, and method 2 requires too much memory. Both cases are problematic: either each new password takes too long to compute, or precomputing all possibilities and storing all the results is too large a task.

Is there some compromise that requires less computing power than method 1 and less memory than required for method 2? Indeed, there is. In 1980 Martin Hellman of Stanford University suggested an approach that was improved in 2003 by Philippe Oechslin of the Swiss Federal Institute of Technology in Lausanne and further refined more recently by Gildas Avoine of the National Institute of Applied Sciences of Rennes (INSA Rennes) in France. It demands less computing power than method 1 in exchange for using a little more memory.

The Beauty of the Rainbow

Here is how it works: First, we need a function R that transforms a fingerprint $h(\mathrm{P})$ into a new password $R(h(\mathrm{P}))$. One might, for instance, consider fingerprints as numbers written in the binary numeral system and

consider passwords as numbers written in the K numeral system, where K is the number of allowable symbols for passwords. Then the function R converts data from the binary numeral system to the K numeral system. For every fingerprint $h(P)$, it computes a new password $R(h(P))$.

Now, with this function R, we can precompute data tables called rainbow tables (so named perhaps because of the multicolored way these tables are depicted).

To generate a data point in this table, we start from a possible password P_0, compute its fingerprint, $h(P_0)$ and then compute a new possible password $R(h(P_0))$, which becomes P_1. Next, we continue this process from P_1. Without storing anything other than P_0, we compute the sequence P_1, P_2,... until the fingerprint starts with 20 zeros; that fingerprint is designated $h(P_n)$. Such a fingerprint occurs only once in about 1,000,000 fingerprints because the result of a hash function is similar to result of a uniform random draw, and 2^{20} is roughly equal to 1,000,000. The password/fingerprint pair $[P_0, h(P_n)]$, containing the fingerprint that starts with 20 zeros is then stored in the table.

A very large number of pairs of this type are computed. Each password/fingerprint pair $[P_0, h(P_n)]$ represents the sequence of passwords P_0, P_1,… P_n and their fingerprints, but the table does not store those intermediate calculations. The table thus lists many password/fingerprint pairs and represents many more (the intermediates, such as P_1 and P_2, that can be derived from the listed pairs). But, of course, there may be gaps: some passwords may be absent from all the chains of calculations.

For a good database with almost no gaps, the memory needed to store the calculated pairs is a million times smaller than that needed for method 2, as described earlier. That is less than four one-terabyte hard disks. Easy. Also, as will be seen, using the table to derive passwords from stolen fingerprints is quite doable.

Let us see how the data stored on the hard disks makes it possible to determine a password in a given space in just a few seconds. Assume that there are no gaps; precomputation of the table takes into account all the

passwords of a designated type—for example, 12-character passwords taken from the 26 letters of the alphabet.

A fingerprint f_0 in a stolen data set can be used to reveal the associated password in the following way. Calculate $h(R(f_0))$ to arrive at a new fingerprint, f_1, then calculate $h(R(f_1))$ to get f_2, and so on, until you get to a fingerprint that begins with 20 zeros: f_m. Then check the table to see which original password, P_0, the fingerprint f_m. is associated with. Based on P_0, calculate the passwords and fingerprints h_1, h_2,... that follow until you inevitably generate the original fingerprint f_0, designated h_k. The password you are looking for is the one that gave rise to h_k—in other words, $R(h_{k-1})$, which is one step earlier in the chain of calculations.

The computation time required is what it takes to look for f_m in the table plus the time needed to compute the sequence of fingerprints from the associated password (h_1, h_2,…, h_k)—which is about a million times shorter than the time needed to compute the table itself. In other words, the time needed is quite reasonable.

Thus, doing a (very long) precomputation and storing only part of the results makes it possible to retrieve any password with a known fingerprint in a reasonable amount of time.

The sequences below represent separate chains of calculations leading from passwords (M_o, N_o,..., Q_o) to fingerprints and other passwords, until the desired fingerprint (and thus the password that precedes it) pops out. (The long dotted line represents may other lines similar to the top two.)

$$M_0 \xrightarrow{h} h(M_0) \xrightarrow{R} M_1 \xrightarrow{h} h(M_1) \xrightarrow{R} M_2 \xrightarrow{h} \dots \xrightarrow{h} h(M_n)$$

$$N_0 \xrightarrow{h} h(N_0) \xrightarrow{R} N_1 \xrightarrow{h} h(N_1) \xrightarrow{R} N_2 \xrightarrow{h} \dots \xrightarrow{h} h(N_p)$$

....

$$Q_0 \xrightarrow{h} h(Q_0) \xrightarrow{R} Q_1 \xrightarrow{h} h(Q_1) \xrightarrow{R} Q_2 \xrightarrow{h} \dots \xrightarrow{h} h(Q_r)$$

To summarize, by knowing the beginning and end of each chain of computations (the only things that are stored during precomputation), a hacker can retrieve any password from a fingerprint. In somewhat simplistic terms, starting from a stolen fingerprint—call it fingerprint X—a hacker would apply the R and *h* functions repeatedly, calculating a series of passwords and fingerprints until reaching a fingerprint with 20 zeros in front of it. The hacker would then look up that final fingerprint in the table (Fingerprint C in the example below) and identify its corresponding password (Password C).

Sample Table Excerpt
Password A—Fingerprint A
Password B—Fingerprint B
Password C—Fingerprint C
Password D—fingerprint D

Next, the hacker would apply the *h* and *R* functions again, beginning with the identified password, continuing on until one of the resulting fingerprints in the chain matches the stolen fingerprint:

Sample Calculation
Password C → fingerprint 1 → password 2-- → fingerprint 2 → password 3…. → password 22-- →**fingerprint 23** [a match to fingerprint X!]

The match (fingerprint 23) would indicate that the previous password (password 22), from which the fingerprint was derived, is the one linked to the stolen fingerprint.

Many computations must be done to establish the first and last column of the rainbow table. By storing only the data in these two columns and by recomputing the chain, hackers can identify any password from its fingerprint.

--Originally published: Scientific American Online April 12, 2019.

Art by the Numbers

by Stephen Ornes

We often regard mathematics with a cold reverence. The discipline is driven by rules and principles that are eternal and stoic. There will never be a countable number of primes, for instance, and the digits of pi will go on forever.

Beneath that certainty, however, lies a sublime attractiveness. A proof or equation can have an elegant, aesthetic effect. Mathematicians who study group theory, for example, analyze rules governing rotations or reflections. Visually, these transformations can appear as intensely beautiful symmetries, such as the radial patterns of snowflakes.

Some mathematicians and artists see a false choice between math and art. They choose not to choose. They ask questions using the language of numbers and group theory and find answers in metal, plastic, wood and computer screen. They weave, and they sketch, and they build. Many of them exchange ideas every year at the international Bridges conference on math and the arts or meet at the biennial Gathering 4 Gardner, named for Martin Gardner, who wrote the celebrated Mathematical Games column in this magazine for 25 years.

Now interest in math art appears to be blooming, shown by an uptick in exhibitions and even academic journals. Roots of the current wave go back to the end of the 20th century, but artists today call on a wider spectrum of mathematical muses and use more modern tools. Here are a few of the most striking works.

Borromean Rings Seifert Surface (2008)
Bathsheba Grossman

For more than a decade Grossman, who lives near Boston, has been using 3-D printing to forge mathematical sculptures out of metal. She delights in symmetries, impossibilities and the division of space. The three outer rings here do not touch one another but are still inextricably interlinked. If you remove one, the other two can separate. It is an ancient form called Borromean rings that is seen today in the logo of the International Mathematical Union.

The rings are members of a mathematical family of link forms, each member characterized by three closed curves with no two physically connected. Their interactions are of particular interest to mathematicians who work in knot theory. The surface bounded by the Borromean rings is called a Seifert surface.

Grossman's sculpture is part knot theory and part puzzle. To highlight the curious swoops of the surface, she used a perforated texture that both plays with light and draws attention to the curious topography.

Buddhabrot (1993)

Melinda Green

In the late 20th century a pattern called the Mandelbrot set took much of the math and art worlds by storm. It was a fractal set named for Benoit B. Mandelbrot, the late French-

American mathematician who was the first to organize fractals into a field worthy of investigation. His 1982 book *The Fractal Geometry of Nature* remains a classic.

The set starts with a point on a complex plane, represented by a two-dimensional graph, and that point is used as the initial value for a particular equation. After making the appropriate calculations, take the new answer and plug it back into the equation. Repeat. If the answers do not get too large—increasing a bit, decreasing a bit—then the initial point is in the set.

Plots of such sets show telltale shapes that repeat as you zoom in or out. But until the 1990s the Mandelbrot set had a standard appearance that made it look like a big bug, with little bugs scattered around its edges and smaller bugs attached to those bugs.

Green, a computer programmer, did not like the "bug body" look. So she hammered out a program that showed more detail about the way certain points hopscotch around the plane. What appeared on her monitor was spooky. "I don't know if I literally pinched myself," she says. The image was a convincing facsimile of the Buddha, and Green revised the code to accentuate different colors. Many mathematicians compare the abstractions of mathematics to spiritual experiences, and Green's "Buddhabrot" invokes that bridge explicitly.

Aurora Australis (2010)
Carlo H. Séquin

In the math art world, Séquin, a computer scientist at the University of California, Berkeley, is known for making hundreds of pieces that give body to heady ideas about surfaces, twists and dimensions. He has produced a veritable zoo of pieces out of wood, metal and plastic.

This piece, he says, was inspired by the celestial light show that plays out in the skies of the Southern Hemisphere: the Aurora Australis, or Southern Lights. The twisting ribbon of the sculpture invokes the turning ribbons of light. In the sculpture, the ribbon changes from flat to curved to flat again and connects to itself. If you trace the sculpture's winding path with your finger, you will visit every part of it and wind up back where you started without lifting your finger. The inside surface is also the outside, which makes it a Möbius strip, the simplest known nonorientable surface, which means that you cannot use concepts such as "front" or "back" or "inside out" with it.

According to Séquin , such visuals are not just captivating; they also provide access to heavy mathematical ideas. "It's a way of getting people who hated math to refocus," he says. "It's a way to see math as much, much more than just rote learning."

Hyperbolic Plane/Pseudosphere (2005)
Daina Taimina

Taimina's adventures in geometric handicrafts began in the 1990s, when the now retired mathematician was teaching a class on hyperbolic geometry, a type of non-Euclidean geometry, at Cornell University. In Euclidean geometry, if you have a line and a point not on the line, there is only one other line that both passes through the point and is parallel to your first line. But in non-Euclidean geometries, there may be many lines that pass through the point and do not intersect the first line. This happens because a hyperbolic plane has constant negative curvature. (The surface of a sphere has constant positive curvature; negative curvature is more like what you would find on a saddle.) As a result, the angles of triangles on hyperbolic planes add up to less than 180 degrees. It is the kind of curvy weirdness that shows up as the frill on the edge of a kale leaf.

Taimina wanted to create tactile models so her students could feel the curvature. Crochet, which she has been practicing almost her entire life, seemed like a good fit. With a crochet hook and yarn, she created a hyperbolic surface using a simple recipe, increasing the number of stitches exponentially. The one shown here takes the form of a pseudosphere, which has negative curvature everywhere.

Since then, Taimina has made dozens of models in an array of colors—the largest weighs about 17 pounds—and can claim invention of "hyperbolic crochet." Her method for creating dazzling blobs has only one basic step. "It's very simple," she says. "Keep constant curvature."

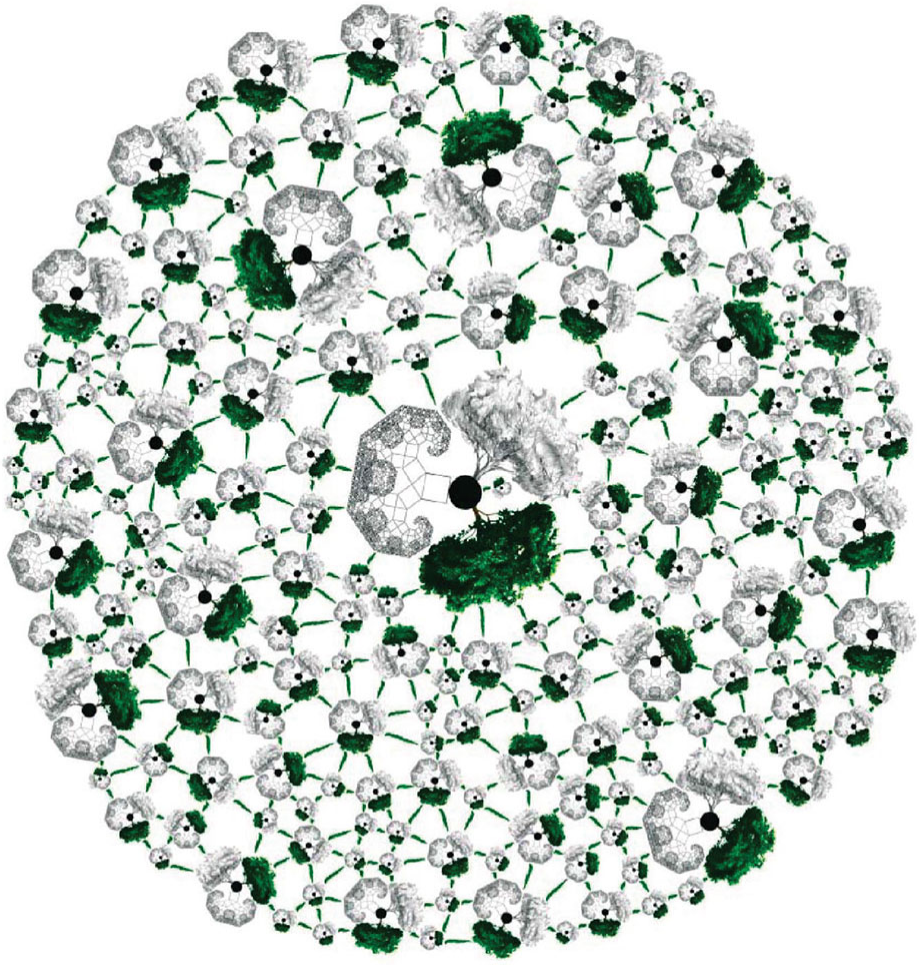
Atomic Tree (2002)
John Sims

Mathematician-artist Sims lives in Sarasota, Fla., and draws inspiration from a range of mathematical ideas. The central image here depicts trees growing on a fractal, which is a pattern that is self-similar: it is the same at every scale, whether you zoom in or out.

Such patterns appear in nature in bushy broccoli crowns and jagged mountain ranges, and scientists have used them to study a range of phenomena, from the structure of the cosmos to the flight patterns of birds.

This figure combines images of a real tree, a drawn tree and a fractal in the shape of a tree. It “speaks to the intersection of math, art and nature,” Sims says. In “Atomic Tree,” the joined shapes serve as building blocks, repeated large and small and connected to form one big network.

Sims first showcased this piece at MathArt/ArtMath, a 2002 exhibition he co-curated at the Ringling College of Art and Design. He has also produced many works inspired by the sequence of digits of pi, including quilts and dresses. With fellow mathematician-artist Vi Hart, in 2015 he produced a “Pi Day Anthem,” in which the duo recites the digits of pi over an infectious drum and bass groove.



Scarabs (2018)

Bjarne Jespersen

Jespersen calls himself a magic wood-carver. The Danish artist aspires to disbelief: he wants people to see, hold and move his wood creations and still not believe in them. “I’m

more of a magician than I am a mathematician or an artist," he says.

If you hold this ball in your hands, you quickly realize that each of these beetles jiggles independently from the rest, and yet they are interlocked and unable to be removed from the whole without breaking something. The ball is carved from a single block of beech.

Jespersen has been inspired by Dutch artist M. C. Escher, much of whose art was mathematical in spirit. Escher popularized tessellations, which are geometric shapes that fit together in a repeated pattern that covers, or tiles, a plane. Mathematicians have long investigated the properties of tessellations—not only of a flat surface but also of higher dimensions. (Escher himself was inspired by the use of tessellations in Islamic art; in particular, the patterns used to decorate the walls of the Alhambra in southern Spain.) Jesper sen's "Scarabs" uses the little bug as the basis for its tessellation.

--Originally published: Scientific American 319(2); 68-73 (August 2018).

SECTION 4

The Outer Limits

Gödel’s Proof

by Ernest Nagel and James R. Newman

In 1931 a young mathematician of 25 named Kurt Gödel published in a German scientific periodical a paper which was read only by a few mathematicians. It bore the forbidding title: "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems." It dealt with a subject that has never attracted more than a small group of investigators, and its reasoning was so novel and complex that it was unintelligible even to most mathematicians. But Gödel's paper has become a landmark of science in the 20th century. As "Gödel's proof," its general conclusions have become known to many scientists, and appreciated to be of revolutionary importance. Gödel's achievement has been recognized by many honors; not long after his paper appeared the young man was invited from Vienna to join the Institute for Advanced Study at Princeton, and he has been a permaanent member of the Institute since 1938. When Harvard University awarded him an honorary degree in 1952, the citation described his proof as one of the most important advances in logic in modern times.

Gödel attacked a central problem in the foundations of mathematics. The axiomatic method invented by the Creeks has always been regarded as the strongest foundation for erecting systems of mathematical thinking. This method, as every student of logic knows, consists in assuming certain propositions or axioms (*e.g.*, if equals be added to equals, the wholes are equal) and deriving other propositions or theorems from the axioms. Until recent times the only branch of mathematics that was considered by most students to be established on sound axiomatic foundations was geometry. But within the past two centuries powerful and rigorous systems of axioms have been developed for other branches of mathematics, including the familiar arithmetic of whole numbers. Mathematicians came to hope and

believe that the whole realm of mathematical reasoning could be brought into order by way of the axiomatic method.

Gödel's paper put an end to this hope. He confronted mathematicians with proof that the axiomatic method has certain inherent limitations which rule out any possibility that even the ordinary arithmetic of whole numbers can ever be fully systematized by its means. What is more, his proofs brought the astounding and melancholy revelation that it is impossible to establish the logical consistency of any complex deductive system except by assuming principles of reasoning whose own internal consistency is as open to question as that of the system itself.

Gödel's paper was not, however, altogether negative. It introduced into the foundations of mathematics a new technique of analysis which is comparable in fertility with Rene Descartes's historic introduction of the algebraic method into geometry. Gödel's work initiated whole new branches of study in mathematical logic. It provoked a reappraisal of mathematical philosophies, and indeed of philosophies of knowledge in general.

His epoch-making paper is still not widely known, and its detailed demonstrations are too complex to be followed by a nonmathematician, but the main outlines of his argument and conclusions can be understood. This article will recount the background of the problem and the substance of Gödel's findings.

The New Mathematics

The 19th century witnessed a tremendous surge forward in mathematical research. Many fundamental problems that had long resisted solution were solved; new areas of mathematical study were created; foundations were newly built or rebuilt for various branches of the discipline. The most revolutionary development was the construction of new geometries by replacing certain of Euclid's axioms with different ones. In particular the modification of Euclid's parallel axiom led to immensely fruitful results. It was this successful departure that stimulated the development of an axiomatic basis for other branches of mathematics which had been cultivated in a more or less intuitive manner. One important conclusion that emerged from this critical examination of the foundations of mathematics was that the traditional conception of mathematics as the "science of

quantity" was inadequate and misleading. For it became evident that mathematics was most essentially concerned with drawing necessary conclusions from a given set of axioms (or postulates). It was thus recognized to be much more "abstract" and "formal" than had been traditionally supposed: more "abstract" because mathematical statements can be construed to be about anything whatsoever, not merely about some inherently circumscribed set of objects or traits of objects; more "formal" because the validity of a mathematical demonstration is grounded in the structure of statements rather than in the nature of a particular subject matter. The postulates of any branch of demonstrative mathematics are not inherently about space, quantity, apples, angles or budgets, and any special meaning that may be associated with the postulates' descriptive terms plays no essential role in the process of deriving theorems. The question that confronts a pure mathematician (as distinct from the scientist who employs mathematics in investigating a special subject matter) is not whether the postulates he assumes or the conclusions he deduces from them are true, but only whether the alleged conclusions are in fact the necessary logical consequences of the initial assumptions. This approach recalls Bertrand Russell's famous epigram: Pure mathematics is the subject in which we do not know what we are talking about, nor whether what we are saying is true.

A land of rigorous abstraction, empty of all familiar landmarks, is certainly not easy to get around in. But it offers compensations in the form of a new freedom of movement and fresh vistas. As mathematics became more abstract, men's minds were emancipated from habitual connotations of language and could construct novel systems of postulates. Formalization led in fact to a great variety of systems of considerable mathematical interest and value. Some of these systems, it must be admitted, did not lend themselves to interpretations as obviously intuitive ("common sense") as those of Euclidean geometry or arithmetic, but this fact caused no alarm. Intuition, for one thing, is an elastic faculty. Our children will have no difficulty in accepting as intuitively obvious the paradoxes of relativity, just as we do not boggle at ideas which were regarded as wholly unintuitive a couple of generations ago. Moreover intuition, as we all know, is not a safe guide: it cannot be used safely as a criterion of either truth or fruitfulness in scientific explorations.

However, the increased abstractness of mathematics also raised a more serious problem. When a set of axioms is taken to be about a definite and familiar domain of objects, it is usually possible to ascertain whether the axioms are indeed true of these objects, and if they are true, they must also be mutually consistent. But the abstract non-Euclidean axioms appeared to be plainly false as descriptions of space, and, for that matter, doubtfully true of anything. Thus the problem of establishing the internal consistency of non-Euclidean systems was formidable. In Riemannian geometry, for example, the famous parallel postulate of Euclid is replaced by the assumption that through a given point outside a line no parallel to the line can be drawn in the same plane. Now suppose the question: Is the Riemannian set of postulates consistent? They are apparently not true of the ordinary space of our experience. How then is their consistency to be tested? How can one prove they will not lead to contradictory theorems?

A general method for solving this problem was proposed. The underlying idea was to find a "model" for the postulates so that each postulate was converted into a true statement about the model. The procedure goes something like this. Let us take the word "class" to signify a collection of distinguishable elements, or "members." (For example, the class of prime numbers less than 10 is a collection consisting of 2, 3, 5 and 7 as members.) Suppose now we consider two purely abstract classes, K and L, concerning which these postulates are given:

1. Any two members of K are contained in just one member of L.

2. No member of K is contained in more than two members of L.

3. The members of K are not all contained in a single member of L.

4. Any two members of L contain just one member of K.

5. No member of L contains more than two members of K.

From this little set we can derive, by using customary rules of inference, certain theorems. For example, it can be shown that K contains just three members. But is the set a consistent one, so that mutually contradictory theorems can never be derived from it? This is where we invoke the help of a model, or interpretation, of the classes. Let K be the vertices of a triangle, and L its sides. Each of the five abstract postulates is then converted into a

true statement: *e.g.*, the first postulate asserts that any two of the vertices are contained on just one side. In this way the set is proved to be consistent.

At first thought such a procedure may seem to suffice to establish the consistency of an abstract system such as plane Riemannian geometry. We may adopt a model embodying the Riemannian postulates in which the expression "plane" signifies the surface of a Euclidean sphere; the expression "point," a point on this surface; the expression "straight line," an arc of a great circle on this surface, and so on. Each Riemannian postulate can then be converted into a theorem of Euclid. For example, on this interpretation the Riemannian parallel postulate reads as follows: Through a point on the surface of a sphere, no arc of a great circle can be drawn parallel to a given arc of a great circle.

Unhappily this method is vulnerable to a serious objection; namely, that it attempts to solve a problem in one domain merely by shifting the problem to another (or, to put it another way, we invoke Euclid to demonstrate the consistency of a system which subverts Euclid). Riemannian geometry is proved to be consistent only if Euclidean geometry is consistent. Query, then: Is Euclidean geometry consistent? If we attempt to answer this question by invoking yet another model, we are no closer to our goal. In short, any proof obtained by this method will be only a "relative" proof of consistency, not an absolute proof.

So long as we can interpret a system by a model containing only a finite number of elements, we have no great difficulty in proving the consistency of its postulates. For example, the triangle model which we used to test the K and L class postulates is finite, and accordingly it is comparatively simple to determine by actual inspection whether the postulates are "true" and hence consistent. Unfortunately most of the postulate systems that constitute the foundations of important branches of mathematics cannot be mirrored in finite models; they can be satisfied only by nonfinite ones. In a well-known set of axioms for elementary arithmetic one of the axioms asserts that every integer in the sequence of whole numbers has an immediate successor which differs from any preceding integer. Obviously any model used to test the set of postulates must mirror the infinity of elements postulated by this axiom. It follows that the truth (and so the

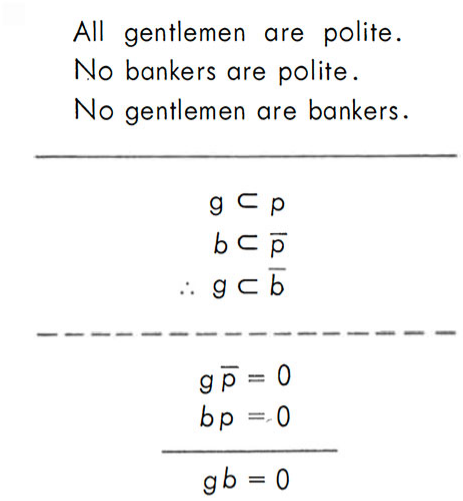
consistency) of the set cannot be established by inspection and enumeration. Apparently we have reached an impasse.

Russell's Paradox

It may be tempting to suggest at this point that we can be sure that a set of postulates is consistent, i.e., free from contradictions, if the basic notions employed are transparently "clear" and "certain." But the history of thought has not dealt kindly with the doctrine of intuitive knowledge implicit in this suggestion. In certain areas of mathematical research radical contradictions have turned up in spite of the "intuitive" clarity of the notions involved in the assumptions, and despite the seemingly consistent character of the intellectual constructions performed. Such contradictions (technically called "antinomies") have emerged, for example, in the theory of infinite numbers developed by Georg Cantor in the 19th century. His theory was built on the elementary and seemingly "clear" concept of class. Since modern systems in other branches of mathematics, particularly elementary arithmetic, have been built on the foundation of the theory of classes, it is pertinent to ask whether they, too, are not infected with contradictions.

In point of fact, Bertrand Russell constructed a contradiction within the framework of elementary logic itself. It is precisely analogous to the contradiction first developed in the Cantorian theory of infinite classes. Russell's antinomy can be stated as follows: All classes apparently may be divided into two groups: those which do not contain themselves as members, and those which do. An example of the first is the class of mathematicians, for patently the class itself is not a mathematician and is therefore not a member of itself. An example of the second is the class of all thinkable concepts, for the class of all thinkable concepts is itself a thinkable concept, and is therefore a member of itself. We shall call the first type of class "normal," and the second type "nonnormal." Now let N stand for the class of all normal classes. We ask whether N itself is a normal class. If so, it is a member of itself. But in that case N is nonnormal, because by definition a class which contains itself is non-normal. Yet if N is non-normal and thus a member of itself, it must be normal, because by definition all the members of N are normal. In short, N is normal if and only if N is non-normal. This fatal contradiction results from an uncritical use of the apparently pellucid notion of class.

Other paradoxes were found later, each of them constructed by means of familiar and seemingly cogent modes of reasoning. Non-finite models by their very nature involve the use of possibly inconsistent sets of postulates. Thus it became clear that, although the model method for establishing the consistency of axioms is an invaluable mathematical tool, that method does not supply a final answer to the problem it was designed to resolve.



SYMBOLIC LOGIC was invented in the middle of the 19th century by the English mathematician George Boole. In this illustration a syllogism is translated into his notation in two different ways. In the upper group of formulas, the symbol $\subset$ means "is contained in." Thus $g \subset p$ says that the class of gentlemen is included in the class of polite persons. In the equations below two letters together mean the class of things having both characteristics. For example, bp means the class of individuals who are bankers and polite. The second equation in the group says that this class has no members. A line above a letter means "not." (Not-p, for example, means impolite.)

Hilbert's Meta-Mathematics

The eminent German mathematician David Hilbert then adopted the opposite approach of eschewing models and draining mathematics of any meaning whatever. In Hilbert's complete formalization, mathematical expressions are regarded simply as empty signs. The postulates and theorems constructed from the system of signs (called a calculus) are simply sequences of meaningless marks which are combined in strict

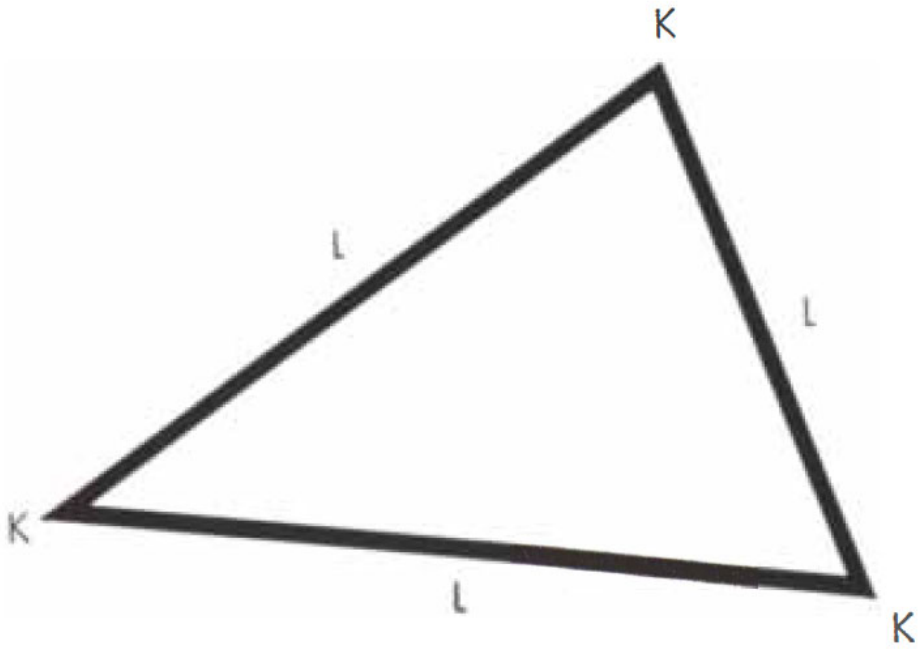
agreement with explicitly stated rules. The derivation of theorems from postulates can be viewed as simply the transformation of one set of such sequences, or "strings," into another set of "strings," in accordance with precise rules of operation. In this manner Hilbert hoped to eliminate the danger of using any unavowed principles of reasoning.

Formalization is a difficult and tricky business, but it serves a valuable purpose. It reveals logical relations in naked clarity, as does a cut-away working model of a machine. One is able to see the structural patterns of various "strings" of signs: how they hang together, how they are combined, how they nest in one another, and so on. A page covered with the "meaningless" marks of such a formalized mathematics does not *assert* anything—it is simply an abstract design or a mosaic possessing a certain structure. But configurations of such a system can be deScribed, and statements can be made about their various relations to one another. One may say that a "string" is pretty, or that it resembles another "string," or that one "string" appears to be made up of three others, and so on. Such statements will evidently be meaningful.

Now it is plain that any meaningful statements about a meaningless system do not themselves belong to that system. Hilbert assigned them to a separate realm which he called "meta-mathematics." Meta-mathematical statements are statements *about* the signs and expressions of a formalized mathematical system; about the kinds and arrangements of such signs when they are combined to form longer strings of marks called "formulas," or about the relations between formulas which may obtain as a consequence of the rules of manipulation that have been specified for them.

A few examples will illustrate Hilbert's distinction between mathematics (a system of meaningless expressions) and meta-mathematics (statements about mathematics). Consider the arithmetical expression 2+3=5. This expression belongs to mathematics and is constructed entirely out of elementary arithmetical signs. Now we may make a statement about the displayed expression, viz.: "'2+3=5' is an arithmetical formula." The statement does not express an arithmetical fact: it belongs to meta-mathematics, because it characterizes the string of arithmetical signs. Similarly the expression x=x belongs to mathematics, but the statement "'x' is a variable" belongs to meta-mathematics. We may also make the

following meta-mathematical statement: "The formula '0=0' is derivable from the formula 'x=x' by substituting the numeral '0' for the variable 'x'." This statement specifies in what manner one arithmetical formula can be obtained from another formula, and thereby describes how the two formulas are related to each other. Again, we may make the meta-mathematical statement: "'0 ≠ 0' is not a theorem." It says that the formula in question is not derivable from the axioms of arithmetic, or in other words, that a certain relation does not hold between the specified formulas of the system. Finally, the following statement also belongs to metamathematics: "Arithmetic is consistent" (i.e., it is not possible to derive from the axioms of arithmetic both the formula 0 = 0 and also the formula 0 ≠ 0).



MODEL for a set of postulates about two classes, K and L, is a triangle whose vertices are the members of K and whose sides are the members of L. The geometrical model shows that the postulates are consistent.

Upon this foundation-separation of meta-mathematical descriptions from mathematics itself-Hilbert attempted to build a method of "absolute" proof of the internal consistency of mathematical systems. Specifically, he sought to develop a theory of proof which would yield demonstrations of consistency by an analysis of the purely structural features of expressions in completely formalized (or "uninterpreted") calculi. Such an analysis consists exclusively of noting the kinds and arrangements of signs in formulas and determining whether a given combination of signs can be obtained from others in accordance with the explicitly stated rules of operation. An absolute proof of the consistency of arithmetic, if one could be constructed, would consist in showing by meta-mathematical procedures of a "finitistic" (non-infinite) character that two "contradictory" formulas, such as (0 = 0) and its negation, cannot both be derived from the axioms or initial formulas by valid rules of inference.

It may be useful, by way of illustration, to compare meta-mathematics as a theory of proof with the theory of chess. Chess is played with 32 pieces of specified design on a square board containing 64 square subdivisions, where the pieces may be moved in accordance with fixed rules. Neither the pieces, nor the squares, nor the positions of the pieces on the board signify anything *outside* the game. In this sense the pieces and their configurations on the board are "meaningless." Thus the game is analogous to a formalized mathematical calculus. The pieces and the squares of the board correspond to the elementary signs of the calculus; the initial positions of the pieces correspond to the axioms or initial formulas of the calculus; their subsequent positions correspond to formulas derived from the axioms (*i.e.*, to the theorems), and the rules of the game correspond to the rules of inference for the calculus. Now, though configurations of pieces on the board are "meaningless," statements about these configurations, like meta-mathematical statements about mathematical formulas, are quite meaningful. A "meta-chess" statement may assert that there are 20 possible opening moves for White, or that, given a certain configuration of pieces on the board with White to move, Black is mate in three moves. Moreover, one can prove general "meta-chess" theorems on the basis of the finite number of permissible configurations on the board. The meta-chess theorem about the number of possible opening moves for White can be established in this way, if White has only two Knights, it is impossible for White to mate Black. These and other "meta-chess" theorems can, in other words, be

proved by finitistic methods of reasoning, consisting in the examination of each of a finite number of configurations that can occur under stated conditions. The aim of Hilbert's theory of proof, similarly, was to demonstrate by such finitistic methods the impossibility of deriving certain contradictory formulas in a calculus.

The *Principia*

It was Hilbert's approach, coupled with the formalization of logic itself in the famous *Principia Mathematica* by Alfred North Whitehead and Bertrand Russell, that led to the crisis to which Gödel supplied a final answer.

The grand object of *Principia*, published in 1910, was to demonstrate that mathematics is only a chapter of logic. But it made two contributions which are of particular interest to us here. First, following up work by the 19th-century pioneer George Boole, it supplied a system of symbols which permitted all statements of pure mathematics to be codified in a standard manner. Secondly, it stated in explicit form most of the rules of formal logic that are employed in mathematical proofs. Thus *Principia* provided an essential instrument for investigating the entire system of arithmetic as a system of "meaningless" marks which could be operated upon in accordance with explicitly stated rules.

We turn now to the formalization of a small portion of *Principia*, namely, the elementary logic of propositions. The task is to convert this fragment into a "meaningless" calculus of uninterpreted signs and to demonstrate a method of proving that the calculus is free from contradictions.

Four steps are involved. First we must specify the complete "vocabulary" of signs to be employed in the calculus. Second, we state the "formation rules" (the rules of "grammar") which indicate the combinations of signs permissible as formulas (or "sentences"). Third, we specify the "transformation rules," which tell how formulas may be derived from others. Finally, we select certain formulas as axioms which serve as foundations for the entire system. The "theorems" of the system are all the formulas, including the axioms, that can be derived from the axioms by applying the transformation rules. A "proof" consists of a finite sequence of

legitimate formulas, each of which is either an axiom or is derivable from preceding formulas in the sequence by the transformation rules.

The vocabulary for the elementary logic of propositions (often also called the "sentential calculus") is extremely simple. The "sentential" variables (which correspond to sentences or statements) are certain letters: *p, q, r*' and so on. Then there are several connectives: ~, which stands for "not"; ∨, which stands for "or"; ⊃, which stands for "if . . . then," and •, which stands for "and." Parenthesis marks are used as signs of punctuation.

Each sentential variable counts as a formula, and the signs may be combined according to the formation rules to form other formulas: *e.g.*, $p \supset q$. If a given sentence $(p \supset q)$ is a formula, so is its negation $\sim (p \supset q)$. If two sentences, S_1 and S_2, are formulas, so is the combination $(S_1) \vee (S_2)$. Similar conventions apply to the other connectives.

For transformations there are just two rules. One, the rule of substitution, says that if a sentence containing sentential variables has been assumed, any formulas may be substituted everywhere for these variables, so that the new sentence will count as a logical consequence of the original one. For example, having accepted $p \supset p$ (if *p*, then *p*), we can always substitute *q* for *p*, obtaining as a theorem the formula $q \supset q$; or we may substitute $(p \vee q)$ for *p*, obtaining $(p \vee q) \supset (p \vee q)$. The other rule, that of detachment, simply says that if the sentences S_1 and $S_1 \supset S_2$ are logically true, we may also accept as logically true the sentence S_2.

The calculus has four axioms, essentially those of *Principia*, which are given in the table at the top of this page, along with nonsensical English sentences to illustrate their independence of meaning. The clumsiness of the translations, especially in the case of the fourth axiom, will perhaps help the reader to realize the advantages of using a special symbolism.

1	$(p \vee p) \supset p$ If either p or p, then p	If either Henry VIII was a boor or Henry VIII was a boor, then Henry VIII was a boor.
2	$p \supset (p \vee q)$ If p, then either p or q	If psychoanalysis is valid then either psychoanalysis is valid or headache powders are better.
3	$(p \vee q) \supset (q \vee p)$ If either p or q, then either q or p	If either Immanuel Kant was punctual or Hollywood is sinful then either Hollywood is sinful or Immanuel Kant was punctual.
4	$(p \supset q) \supset [(r \vee p) \supset (r \vee q)]$ If p implies q, then (either r or p) implies (either r or q)	If ducks waddle implies that $\sqrt{2}$ is a number then (either Churchill drinks brandy or ducks waddle) implies (either Churchill drinks brandy or $\sqrt{2}$ is a number).

SENTENTIAL CALCULUS, or the elementary logic of propositions, is based on four axioms. The nonsense statements illustrate how general is the "meaning" of the symbols.

Search for a Proof

Each of these axioms may seem "obvious" and trivial. Nevertheless it is possible to derive from them with the help of the stated transformation rules an indefinitely large class of theorems which are far from obvious or trivial. However, at this point we are interested not in deriving theorems from the axioms but in showing that this set of axioms is not contradictory, We wish to prove that, using the transformation rules, it is impossible to derive from the axioms any formula S (*i.e.*, any expression which would normally count as a sentence) together with its negation ~ S,

Now it can be shown that p $\supset$ (~ $p \supset q$) (if p, then if not-p then q) is a theorem in the calculus. Let us suppose, for the sake of demonstration, that a formula S and its contradictory ~ S were both deducible from the axioms, and test the consequences by means of this theorem. By substituting S for p

in the theorem, as permitted by the rule of substitution, we first obtain S ⊃ (~ S ⊃ *q*). From this, assuming S to be demonstrably true, we could next obtain, by the detachment rule, ~ S ⊃ *q*. Finally, if we assume ~ S also is demonstrable, by the detachment rule we would get *q*. Since we can substitute any formula whatsoever for *q*, this means that any formula whatsoever would be deducible from the axioms. Thus if both S and its contradictory ~ S were deducible from the axioms, then *any* formula would be deducible. We arrive, then, at the conclusion that if the calculus is not consistent (*i.e.*, if both S and ~ S are deducible) any theorem can be derived from the axioms. Accordingly, to prove the consistency of the calculus, our task is reduced to finding at least one formula which cannot be derived from the axioms.

The way this is done is to employ meta-mathematical reasoning upon the system before us. The actual procedure is elegant. It consists in finding a characteristic of formulas which satisfies the three following conditions. (1) it is common to all four axioms; (2) it is "hereditary," that is, any formula derived from the axioms (*i.e.*, any theorem) must also have the property; (3) there must be at least one formula which does not have the characteristic and is therefore not a theorem. If we succeed in this threefold task, we shall have an absolute proof of the consistency of the axioms. If we can find an array of signs that conforms to the requirements of being a formula but does not possess the specified characteristic, this formula cannot be a theorem. In other words, the finding of a single formula which is not a theorem suffices to establish the consistency of the system.

Let us choose as a characteristic of the required kind the property of being a "tautology." In common parlance a tautology is usually considered to be a redundant statement such as: "John is the father of Charles and Charles is a son of John." But in logic a tautology is defined as a statement which excludes no logical possibilities—*e.g.*, "Either it is raining or it is not raining." Another way of putting this is to say that a tautology is "true in all possible worlds." We apply this definition to formulas in the system we are considering. A formula is said to be a tautology if it is invariably true regardless of whether its elementary constituents (*p*, *q*, *r* and so on) are true or false. Now all four of our axioms plainly possess the property of being tautologous. For example, the first axiom, (*p* ∨ *p*) ⊃ *p*, is true regardless of whether *p* is assumed to be true or is assumed to be fake. The axiom says,

for instance: "If either Mount Rainier is 20,000 feet high or Mount Rainier is 20,000 feet high, then Mount Rainier is 20,000 feet high." It makes no difference whether Mount Rainier is actually 20,000 feet high or not: the statement is still true in either case. A similar demonstration can be made for the other axioms.

Next it is possible to prove that the property of being a tautology is hereditary under the transformation rules, though, we shall not turn aside to give the demonstration. It follows that every formula properly derived from the axioms (*i.e.*, every theorem) must be a tautology. Having performed these two steps, we are ready to look for a formula which does not possess the characteristic of being a tautology. We do not have to look very hard. For example, $p \vee q$ fits the requirements. Clearly it is not a tautology; it is the same as saying: "Either John is a philosopher or Charles reads *Scientific American*." This is patently not a truth of logic; it is not a sentence that is true irrespective of the truth or falsity of its elementary constituents. Thus $p \vee q$, though it purports to be a gosling, is in fact a duckling; it is a formula but it is not a theorem.

We have achieved our goal. We have found at least one formula which is not a theorem, therefore the axioms must be consistent.

Gödel’s Answer

The sentential calculus is an example of a mathematical system for which the objectives of Hilbert's theory of proof are fully realized. But this calculus codifies only a fragment of formal logic. The question remains: Can a formalized system embracing the whole of arithmetic be proved consistent in the sense of Hilbert's program?

This was the conundrum that Gödel answered. His paper in 1931 showed that all such efforts to prove arithmetic to be free from contradictions are doomed to failure.

His main conclusions were twofold. In the first place, he showed that it is impossible to establish a meta-mathematical proof of the consistency of a system comprehensive enough to contain the whole of arithmetic-unless, indeed, this proof itself employs rules of inference much more powerful than the transformation rules used in deriving theorems within the system. In short, one dragon is slain only to create another.

Gödel's second main conclusion was even more surprising and revolutionary, for it made evident a fundamental limitation in the power of the axiomatic method itself. Gödel showed that *Principia*, or any other system within which arithmetic can be developed, is essentially incomplete. In other words, given any consistent set of arithmetical axioms, there are true arithmetical statements which are not derivable from the set. A classic illustration of a mathematical "theorem" which has thwarted all attempts at proof is that of Christian Goldbach, stating that every even number is the sum of two primes. No even number has ever been found which is not the sum of two primes, yet no one has succeeded in finding a proof that the rule applies without exception to all even numbers. In reply to Gödel it might be suggested that the set of arithmetical axioms could be modified or expanded to make "underivable" statements derivable. But Gödel showed that this approach promises no final cure. That is, even if any finite number of other axioms is added, there will always be further arithmetical truths which are not formally derivable.

How did Gödel prove his conclusions? His paper is difficult. A reader must master 46 preliminary definitions, together with several important preliminary theorems, before he gets to the main results. We shall take a much easier road; nevertheless we hope at least to offer glimpses of the argument.

Gödel Numbers

Gödel first devised a method of assigning a number as a label for each elementary sign, each formula and each proof in a formalized system. To the elementary signs he attached as "Gödel numbers" the integers from 1 to 10; to the variables he assigned numbers according to certain rules.

CONNECTIVES AND ELEMENTARY SIGNS

SIGNS	GÖDEL NUMBER	MEANING
~	1	not
v	2	or
⊃	3	If . . . then
∃	4	There is an . . .
=	5	equals
0	6	zero
S	7	The next following number
(	8	punctuation mark
)	9	punctuation mark
,	10	punctuation mark

SENTENTIAL VARIABLES (EACH DESIGNATED BY A NUMBER GREATER THAN 10 AND DIVISIBLE BY 3)

VARIABLES	GÖDEL NUMBER	SAMPLE
p	12	Henry VIII was a boor.
q	15	Headache powders are better.
r	18	Ducks waddle.
etc.		

INDIVIDUAL VARIABLES (EACH DESIGNATED BY A NUMBER GREATER THAN 10 WHICH LEAVES A REMAINDER OF 1 WHEN DIVIDED BY 3)

VARIABLES	GÖDEL NUMBER	MEANING
x	13	a numerical variable
y	16	a numerical variable
z	19	a numerical variable
etc.		

PREDICATE VARIABLES (EACH DESIGNATED BY A NUMBER GREATER THAN 10 WHICH LEAVES A REMAINDER OF 2 WHEN DIVIDED BY 3)

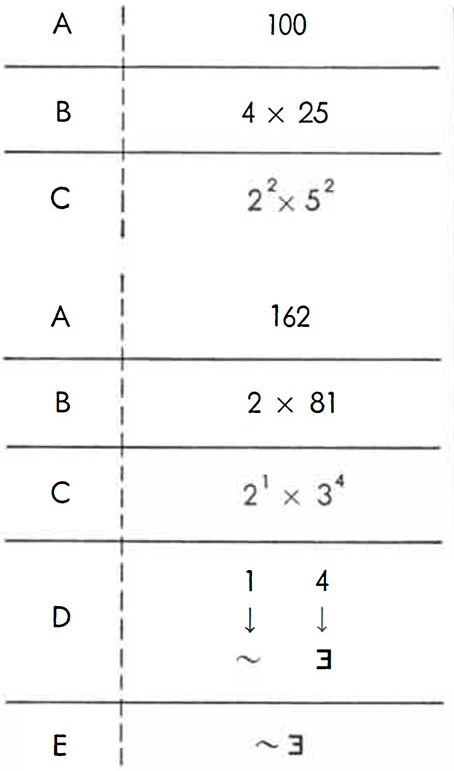
VARIABLES	GÖDEL NUMBER	SAMPLE
P	14	Being a boor
Q	17	Being a headache powder
R	20	Being a duck
etc.		

ELEMENTARY GÖDEL NUMBERS are assigned to every symbol used in his system of symbolic logic in accordance with the orderly scheme which is illustrated in the table above.

To see how a number is given to a formula of the system, let us take this formula: (∃ *x*) (*x* = S*y*), which reads literally "there is an *x*, such that *x* is the immediate successor of *y*" and in effect says that every number has an immediate successor. The numbers associated with the formula's 10 successive signs are, respectively, 8, 4, 13, 9, 8, 13, 5, 7, 16, 9. Now these numbers are to be used as exponents, or powers, of the first 10 prime numbers (*i.e.*, 2, 3, 5 and so on). The prime numbers, raised to these powers, are multiplied together. Thus we get the number $2^8 \times 3^4 \times 5^{13} \times 7^9 \times 11^8 \times 13^{13} \times 17^5 \times 19^7 \times 23^{16} \times 29^9$. The product is the Gödel number of the formula. In the same way every formula can be represented by a single unique number.

We can assign a number to a sequence of formulas, such as may occur in some proof, by a similar process. Let us say that we have a sequence of two

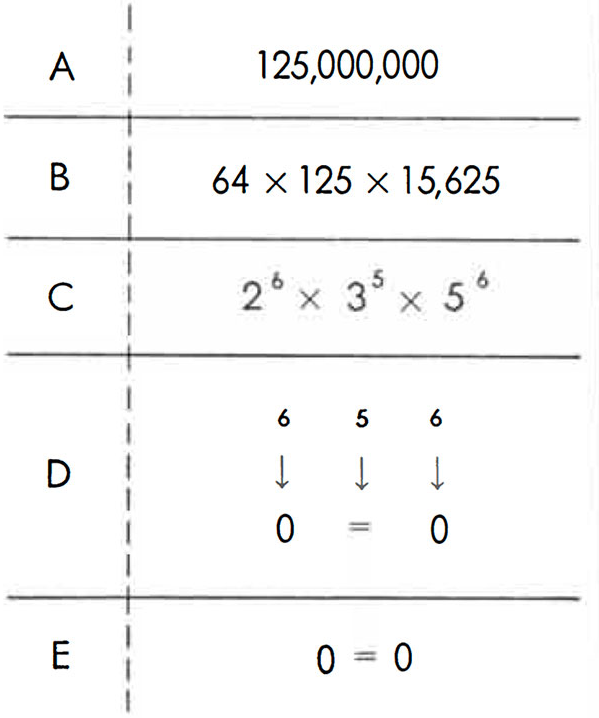
formulas, the second derived from the first. For example, by substituting 0 for *y* in the formula given above, we derive (∃ *x*) (*x*=S0), which says that 0 has an immediate successor. Now the first and second formulas are identified by Gödel numbers which we shall call *m* and *n*, respectively. To label this sequence, we use the Gödel numbers *m* and *n* as exponents and multiply the first two primes (2 and 3) raised to these powers. That is to say, the Gödel number that identifies the sequence is 2^m X 3^n. In like manner we can give a number to any sequence of formulas or any other expression in the system.



GÖDEL NUMBERS of formulas are constructed by raising the prime numbers, in sequence, to powers which are the Gödel numbers of the symbols involved. Thus 100 is not a Gödel number because its factors skip the prime number 3. On the other hand, 162 is the Gödel number for "there is not."

What has been done so far is to establish a method for completely arithmetizing a formal system. The method is essentially a set of directions for making a one-to-one correspondence between specific numbers and the various elements or combinations of elements of the system. Once an

expression is given, it can be uniquely numbered. But more than that, we can retranslate any Gödel number into the expression it represents by factoring it into its component prime numbers, which can be done in only one way, as we know from a famous theorem of arithmetic. In other words, we can take the number apart as if it were a machine, see how it was constructed and what went into it, and we can dissect an expression or a proof in the same way.



ARITHMETICAL FORMULA "zero equals zero" has the Gödel number 125 million. Reading down from A to E, the illustration shows how the number is translated into the expression it represents; reading up, how the number for the formula is derived.

This leads to the next step. It occurred to Gödel that meta-mathematical statements can be translated into arithmetical terms by a process analogous to mapping. In geography the spatial relations between points on the spherical earth can be projected onto a flat map; in mathematical physics relations between the properties of electric currents can be mapped in terms of the flow of fluids; in mathematics itself relations in geometry can be

translated into algebra. Gödel saw that if complicated meta-mathematical statements about a system could be translated into, or mirrored by, arithmetical statements within the system itself, an important gain would be achieved in clarity of expression and facility of analysis. Plainly it would be easier to deal with arithmetical counterparts of complex logical relations than with the logical relations themselves. To cite a trivial analogy: If customers in a supermarket are given tickets with numbers determining the order in which they are to be waited on, it is a simple matter to discover, merely by scrutinizing the numbers, how many persons have been served, how many are waiting, who precedes whom and by how many customers, and so on.

What Gödel aimed at was nothing less than the complete arithmetization of meta-mathematics. If each meta-mathematical statement could be uniquely represented in the formal system by a formula expressing a relation between numbers, questions of logical dependence between meta-mathematical statements could be explored by examining the corresponding relations between integers. Gödel did in fact succeed brilliantly in mapping the meta-mathematics of arithmetic upon arithmetic itself. We need cite only one illustration of how a meta-mathematical statement can be made to correspond to a formula in the formal arithmetical system. Let us take the formula $(p \vee p) \supset p$. We may make the meta-mathematical statement that the formula $(p \vee p)$ is the initial part of this formula. Now we can represent this meta-mathematical statement by an arithmetical formula which says in effect that the Gödel number of the initial part is a factor of the Gödel number of the complete formula. Evidently this is so, for the Gödel number of $(p \vee p)$ is $2^8 \times 3^{12} \times 5^2 \times 7^{12} \times 11^9$, while the Gödel number of $(p \vee p) \supset p$ is $2^8 \times 3^{12} \times 5^2 \times 7^{12} \times 11^9 \times 13^3 \times 17^{12}$.

The Undecidable Proposition

We have now arrived at the very heart of Gödel's analysis. He showed how to construct an arithmetical formula, whose Gödel number we shall suppose is *h*, which corresponds to the meta-mathematical statement, *viz*.: "The formula with Gödel number *h* is not demonstrable." In other words, this formula (call it G) in effect asserts its own indemonstrability, though it is a legitimate formula belonging to the formal system of arithmetic. Gödel then proceeded to examine the question whether G is or is not a

demonstrable formula of arithmetic. He was able to show that G is demonstrable if, and only if, its negation, ~ G, also is demonstrable. But if a formula and its negation are both derivable from a set of axioms, obviously the axioms are not consistent. It follows that if arithmetic is consistent, neither G nor its negation is demonstrable. That is to say, G is an undecidable formula of arithmetic. Now from this Gödel proved the indemonstrability of the proposition that arithmetic is consistent. It can be shown that a metamathematical statement of arithmetic's consistency corresponds to a certain arithmetical formula, A, and that the arithmetical formula A ⊃ G (if A, then G) is demonstrable. Thus if A were demonstrable, G would be also. But we have just seen that G is not demonstrable. It follows that A is undecidable. In short, the consistency of arithmetic is undecidable by any meta-mathematical reasoning which can be represented within the formalism of arithmetic.

Gödel's analysis does not exclude a meta-mathematical demonstration of the consistency of arithmetic; indeed, such proofs have been constructed, notably by Gerhard Gentzen, a member of the Hilbert school. But these "proofs" are in a sense pointless, because they employ rules of inference whose own internal consistency is as much open to doubt as is the formal consistency of arithmetic itself. Gentzen's proof employs a rule of inference which in effect permits a formula to be derived from an infinite class of premises. And the employment of this non-finitistic meta-mathematical notion raises once more the difficulty which Hilbert's original program was intended to resolve.

There is another surprise coming. Although the formula G is undecidable, it can be shown by meta-mathematical reasoning that G is nevertheless a *true* arithmetical statement and expresses a property of the arithmetical integers. The argument for this conclusion is quite simple. We need recall only that Gödel mapped meta-mathematical statements upon arithmetical formulas in such a way that every true meta-mathematical statement corresponds to a true arithmetical formula. Now C corresponds to a meta-mathematical statement ("the formula with Gödel number *h* is not demonstrable") which, as we have seen, is true, unless arithmetic is inconsistent. It follows that C itself must be true. We have thus established an *arithmetical* truth by a *meta-mathematical* argument.

So we come to the finale of Gödel's amazing and profound intellectual symphony. Arithmetic is incomplete, in the transparent sense that there is at least one arithmetical truth which cannot be derived from the arithmetical axioms and yet can be established by a metamathematical argument outside the system. Moreover, arithmetic is *essentially* incomplete, for even if the true formula C were taken as an axiom and added to the original axioms, the augmented system would still not suffice to yield formally all the truths of arithmetic: we could still construct a true formula which would not be formally demonstrable within the system. And such would be the case no matter how often we repeated the process of adding axioms to the initial set.

This remarkable conclusion makes evident an inherent limitation in the axiomatic method. Contrary to previous assumptions, the vast "continent" of arithmetical truth cannot be brought into systematic order by way of specifying once for all a fixed set of axioms from which all true arithmetical statements would be formally derivable.

Men and Calculating Machines

The far-reaching import of Gödel's conclusions has not yet been fully fathomed. They show that the hope of finding an absolute proof of consistency for any deductive system expressing the whole of arithmetic cannot be realized, if such a proof must satisfy the finitistic requirements of Hilbert's original program. They also show that there is an endless number of true arithmetical statements which cannot be formally deduced from any specified set of axioms in accordance with a closed set of rules of inference. It follows that an axiomatic approach to the theory of numbers, for example, cannot exhaust the domain of arithmetic truth. Whether an all-inclusive general definition of mathematical or logical truth can be devised, and whether, as Gödel himself appears to believe, only a thoroughgoing Platonic realism can supply such a definition, are problems still under debate.

Gödel's conclusions have a bearing on the question whether a calculating machine can be constructed that would equal the human brain in mathematical reasoning. Present calculating machines have a fixed set of directives built into them, and they operate in a step-by-step manner. But in the light of Gödel's incompleteness theorem, there is an endless set of problems in elementary number theory for which such machines are

inherently incapable of supplying answers, however complex their built-in mechanisms may be and however rapid their operations. The human brain may, to be sure, have built-in limitations of its own, and there may be mathematical problems which it is incapable of solving. But even so, the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines. There is no immediate prospect of replacing the human mind by robots.

Gödel's proof should not be construed as an invitation to despair. The discovery that there are arithmetical truths which cannot be demonstrated formally does not mean that there are truths which are forever incapable of becoming known, or that a mystic intuition must replace cogent proof. It does mean that the resources of the human intellect have not been, and cannot be, fully formalized, and that new principles of demonstration forever await invention and discovery. We have seen that mathematical propositions which cannot be established by formal deduction from a given set of axioms may nevertheless be established by "informal" meta-mathematical reasoning.

Nor does the fact that it is impossible to construct a calculating machine equivalent to the human brain necessarily mean that we cannot hope to explain living matter and human reason in physical and chemical terms. The possibility of such explanations is neither precluded nor affirmed by Gödel's incompleteness theorem. The theorem does indicate that the structure and power of the human mind are far more complex and subtle than any non-living machine yet envisaged. Gödel's own work is a remarkable example of such complexity and subtlety. It is an occasion not for discouragement but for a renewed appreciation of the powers of creative reason.

--Originally published: Scientific American 194(6); 71-90 (June 1956).

The Limits of Reason

by Gregory Chaitin

In 1956 *Scientific American* published an article by Ernest Nagel and James R. Newman entitled “Gödel’s Proof.” Two years later the writers published a book with the same title—a wonderful work that is still in print. I was a child, not even a teenager, and I was obsessed by this little book. I remember the thrill of discovering it in the New York Public Library. I used to carry it around with me and try to explain it to other children.

It fascinated me because Kurt Gödel used mathematics to show that mathematics itself has limitations. Gödel refuted the position of David Hilbert, who about a century ago declared that there was a theory of everything for math, a finite set of principles from which one could mindlessly deduce all mathematical truths by tediously following the rules of symbolic logic. But Gödel demonstrated that mathematics contains true statements that cannot be proved that way. His result is based on two self-referential paradoxes: “This statement is false” and “This statement is unprovable.”

My attempt to understand Gödel’s proof took over my life, and now half a century later I have published a little book of my own. In some respects, it is my own version of Nagel and Newman’s book, but it does not focus on Gödel’s proof. The only things the two books have in common are their small size and their goal of critiquing mathematical methods.

Unlike Gödel’s approach, mine is based on measuring information and showing that some mathematical facts cannot be compressed into a theory because they are too complicated. This new approach suggests that what Gödel discovered was just the tip of the iceberg: an infinite number of true mathematical theorems exist that cannot be proved from any finite system of axioms.

Complexity and Scientific Laws

My story begins in 1686 with Gottfried W. Leibniz's philosophical essay *Discours de métaphysique (Discourse on Metaphysics)*, in which he discusses how one can distinguish between facts that can be described by some law and those that are lawless, irregular facts. Leibniz's very simple and profound idea appears in section VI of the *Discours*, in which he essentially states that a theory has to be simpler than the data it explains, otherwise it does not explain anything. The concept of a law becomes vacuous if arbitrarily high mathematical complexity is permitted, because then one can always construct a law no matter how random and patternless the data really are. Conversely, if the only law that describes some data is an extremely complicated one, then the data are actually lawless.

Today the notions of complexity and simplicity are put in precise quantitative terms by a modern branch of mathematics called algorithmic information theory. Ordinary information theory quantifies information by asking how many bits are needed to encode the information. For example, it takes one bit to encode a single yes/no answer. Algorithmic information, in contrast, is defined by asking what size computer program is necessary to generate the data. The minimum number of bits—what size string of zeros and ones—needed to store the program is called the algorithmic information content of the data. Thus, the infinite sequence of numbers 1, 2, 3, . . . has very little algorithmic information; a very short computer program can generate all those numbers. It does not matter how long the program must take to do the computation or how much memory it must use —just the length of the program in bits counts. (I gloss over the question of what programming language is used to write the program—for a rigorous definition, the language would have to be specified precisely. Different programming languages would result in somewhat different values of algorithmic information content.)

To take another example, the number pi, 3.14159. . . , also has only a little algorithmic information content, because a relatively short algorithm can be programmed into a computer to compute digit after digit. In contrast, a random number with a mere million digits, say 1.341285. . .64, has a much larger amount of algorithmic information. Because the number lacks a

defining pattern, the shortest program for outputting it will be about as long as the number itself:

```
Begin
Print “1.341285. . .64”
End
```

(All the digits represented by the ellipsis are included in the program.) No smaller program can calculate that sequence of digits. In other words, such digit streams are incompressible, they have no redundancy; the best that one can do is transmit them directly. They are called irreducible or algorithmically random.

How do such ideas relate to scientific laws and facts? The basic insight is a software view of science: a scientific theory is like a computer program that predicts our observations, the experimental data. Two fundamental principles inform this viewpoint. First, as William of Occam noted, given two theories that explain the data, the simpler theory is to be preferred (Occam’s razor). That is, the smallest program that calculates the observations is the best theory. Second is Leibniz’s insight, cast in modern terms—if a theory is the same size in bits as the data it explains, then it is worthless, because even the most random of data has a theory of that size. A useful theory is a compression of the data; comprehension is compression. You compress things into computer programs, into concise algorithmic descriptions. The simpler the theory, the better you understand something.

Sufficient Reason

Despite living 250 years before the invention of the computer program, Leibniz came very close to the modern idea of algorithmic information. He had all the key elements. He just never connected them. He knew that everything can be represented with binary information, he built one of the first calculating machines, he appreciated the power of computation, and he discussed complexity and randomness.

If Leibniz had put all this together, he might have questioned one of the key pillars of his philosophy, namely, the principle of sufficient reason—that everything happens for a reason. Furthermore, if something is true, it must be true for a reason. That may be hard to believe sometimes, in the

confusion and chaos of daily life, in the contingent ebb and flow of human history. But even if we cannot always see a reason (perhaps because the chain of reasoning is long and subtle), Leibniz asserted, God can see the reason. It is there! In that, he agreed with the ancient Greeks, who originated the idea.

Mathematicians certainly believe in reason and in Leibniz's principle of sufficient reason, because they always try to prove everything. No matter how much evidence there is for a theorem, such as millions of demonstrated examples, mathematicians demand a proof of the general case. Nothing less will satisfy them.

And here is where the concept of algorithmic information can make its surprising contribution to the philosophical discussion of the origins and limits of knowledge. It reveals that certain mathematical facts are true for no reason, a discovery that flies in the face of the principle of sufficient reason.

Indeed, as I will show later, it turns out that an infinite number of mathematical facts are irreducible, which means no theory explains why they are true. These facts are not just computationally irreducible, they are logically irreducible. The only way to "prove" such facts is to assume them directly as new axioms, without using reasoning at all.

The concept of an "axiom" is closely related to the idea of logical irreducibility. Axioms are mathematical facts that we take as self-evident and do not try to prove from simpler principles. All formal mathematical theories start with axioms and then deduce the consequences of these axioms, which are called theorems. That is how Euclid did things in Alexandria two millennia ago, and his treatise on geometry is the classical model for mathematical exposition.

In ancient Greece, if you wanted to convince your fellow citizens to vote with you on some issue, you had to reason with them—which I guess is how the Greeks came up with the idea that in mathematics you have to prove things rather than just discover them experimentally. In contrast, previous cultures in Mesopotamia and Egypt apparently relied on experiment. Using reason has certainly been an extremely fruitful approach, leading to modern mathematics and mathematical physics and all that goes

with them, including the technology for building that highly logical and mathematical machine, the computer.

So am I saying that this approach that science and mathematics has been following for more than two millennia crashes and burns? Yes, in a sense I am. My counterexample illustrating the limited power of logic and reason, my source of an infinite stream of unprovable mathematical facts, is the number that I call omega.

The Number Omega

The first step on the road to omega came in a famous paper published precisely 250 years after Leibniz's essay. In a 1936 issue of the *Proceedings of the London Mathematical Society*, Alan M. Turing began the computer age by presenting a mathematical model of a simple, general-purpose, programmable digital computer. He then asked, Can we determine whether or not a computer program will ever halt? This is Turing's famous halting problem.

Of course, by running a program you can eventually discover that it halts, if it halts. The problem, and it is an extremely fundamental one, is to decide when to give up on a program that does not halt. A great many special cases can be solved, but Turing showed that a general solution is impossible. No algorithm, no mathematical theory, can ever tell us which programs will halt and which will not. By the way, when I say "program," in modern terms I mean the concatenation of the computer program and the data to be read in by the program.

The next step on the path to the number omega is to consider the ensemble of all possible programs. Does a program chosen at random ever halt? The probability of having that happen is my omega number. First, I must specify how to pick a program at random. A program is simply a series of bits, so flip a coin to determine the value of each bit. How many bits long should the program be? Keep flipping the coin so long as the computer is asking for another bit of input. Omega is just the probability that the machine will eventually come to a halt when supplied with a stream of random bits in this fashion. (The precise numerical value of omega depends on the choice of computer programming language, but omega's

surprising properties are not affected by this choice. And once you have chosen a language, omega has a definite value, just like pi or the number 3.)

Being a probability, omega has to be greater than 0 and less than 1, because some programs halt and some do not. Imagine writing omega out in binary. You would get something like 0.1110100. . . . These bits after the decimal point form an irreducible stream of bits. They are our irreducible mathematical facts (each fact being whether the bit is a 0 or a 1).

Omega can be defined as an infinite sum, and each *N*-bit program that halts contributes precisely $½^N$ to the sum. In other words, each *N*-bit program that halts adds a 1 to the *N*th bit in the binary expansion of omega. Add up all the bits for all programs that halt, and you would get the precise value of omega. This description may make it sound like you can calculate omega accurately, just as if it were the square root of 2 or the number pi. Not so—omega is perfectly well defined and it is a specific number, but it is impossible to compute in its entirety.

How Omega Is Defined

To see how the value of the number omega is defined, look at a simplified example. Suppose that the computer we are dealing with has only three programs that halt, and they are the bit strings 110, 11100 and 11110. These programs are, respectively, 3, 5 and 5 bits in size. If we are choosing programs at random by flipping a coin for each bit, the probability of getting each of them by chance is precisely $½^3$, $½^5$ and $½^5$, because each particular bit has probability ½. So the value of omega (the halting probability) for this particular computer is given by the equation:

omega = $½^3 + ½^5 + ½^5$ = .001 + .00001 + .00001 = .00110

This binary number is the probability of getting one of the three halting programs by chance. Thus, it is the probability that our computer will halt. Note that because program 110 halts we do not consider any programs that start with 110 and are larger than three bits —for example, we do not consider 1100 or 1101. That is, we do not add terms of .0001 to the sum for each of those programs. We regard all the longer programs, 1100 and so on, as being included in the halting of 110. Another way of saying this is that the programs are self-delimiting; when they halt, they stop asking for more bits.

—G.C.

Why Is Omega Incompressible?

I wish to demonstrate that omega is incompressible—that one cannot use a program substantially shorter than *N* bits long to compute the first *N* bits of omega. The demonstration will involve a careful combination of facts about omega and the Turing halting problem that it is so intimately related to. Specifically, I will use the fact that the halting problem for programs up to length *N* bits cannot be solved by a program that is itself shorter than *N* bits.

My strategy for demonstrating that omega is incompressible is to show that having the first *N* bits of omega would tell me how to solve the Turing halting problem for programs up to length *N* bits. It follows from that conclusion that no program shorter than *N* bits can compute the first *N* bits of omega. (If such a program existed, I could use it to compute the first *N* bits of omega and then use those bits to solve Turing's problem up to *N* bits—a task that is impossible for such a short program.)

Now let us see how knowing *N* bits of omega would enable me to solve the halting problem—to determine which programs halt—for all programs up to *N* bits in size. Do this by performing a computation in stages. Use the integer *K* to label which stage we are at: *K* = 1, 2, 3, . . .

At stage K, run every program up to *K* bits in size for K seconds. Then compute a halting probability, which we will call omegaK, based on all the programs that halt by stage K. OmegaK will be less than omega because it is based on only a subset of all the programs that halt eventually, whereas omega is based on all such programs.

As *K* increases, the value of omegaK will get closer and closer to the actual value of omega. As it gets closer to omega's actual value, more and more of omegaK's first bits will be correct —that is, the same as the corresponding bits of omega.

And as soon as the first *N* bits are correct, you know that you have encountered every program up to *N* bits in size that will ever halt. (If there were another such *N*-bit program, at some later-stage *K* that program would halt, which would increase the value of omegaK to be greater than omega, which is impossible.)

So we can use the first *N* bits of omega to solve the halting problem for all programs up to *N* bits in size. Now suppose we could compute the first *N* bits of omega with a program substantially shorter than *N* bits long. We could then combine that program with the one for carrying out the omegaK algorithm, to produce a program shorter than *N* bits that solves the Turing halting problem up to programs of length *N* bits.

But, as stated up front, we know that no such program exists. Consequently, the first *N* bits of omega must require a program that is almost *N* bits long to compute them. That is good enough to call omega incompressible or irreducible. (A compression from *N* bits to almost *N* bits is not significant for large *N*.)

—G.C.

We can be sure that omega cannot be computed because knowing omega would let us solve Turing's halting problem, but we know that this problem is unsolvable. More specifically, knowing the first *N* bits of omega would enable you to decide whether or not each program up to *N* bits in size ever halts. From this it follows that you need at least an *N*-bit program to calculate *N* bits of omega.

Note that I am not saying that it is impossible to compute some digits of omega. For example, if we knew that computer programs 0, 10 and 110 all halt, then we would know that the first digits of omega were 0.111. The point is that the first *N* digits of omega cannot be computed using a program significantly shorter than *N* bits long.

Most important, omega supplies us with an infinite number of these irreducible bits. Given any finite program, no matter how many billions of bits long, we have an infinite number of bits that the program cannot compute. Given any finite set of axioms, we have an infinite number of truths that are unprovable in that system.

Because omega is irreducible, we can immediately conclude that a theory of everything for all of mathematics cannot exist. An infinite number of bits of omega constitute mathematical facts (whether each bit is a 0 or a 1) that cannot be derived from any principles simpler than the string of bits itself. Mathematics therefore has infinite complexity, whereas any individual theory of everything would have only finite complexity and could not capture all the richness of the full world of mathematical truth.

This conclusion does not mean that proofs are no good, and I am certainly not against reason. Just because some things are irreducible does not mean we should give up using reasoning. Irreducible principles—axioms—have always been a part of mathematics. Omega just shows that a lot more of them are out there than people suspected.

So perhaps mathematicians should not try to prove everything. Sometimes they should just add new axioms. That is what you have got to do if you are faced with irreducible facts. The problem is realizing that they are irreducible! In a way, saying something is irreducible is giving up, saying that it cannot ever be proved. Mathematicians would rather die than do that, in sharp contrast with their physicist colleagues, who are happy to be pragmatic and to use plausible reasoning instead of rigorous proof. Physicists are willing to add new principles, new scientific laws, to understand new domains of experience. This raises what I think is an extremely interesting question: Is mathematics like physics?

Mathematics and Physics

The traditional view is that mathematics and physics are quite different. Physics describes the universe and depends on experiment and observation. The particular laws that govern our universe—whether Newton’s laws of motion or the Standard Model of particle physics—must be determined empirically and then asserted like axioms that cannot be logically proved, merely verified.

Mathematics, in contrast, is somehow independent of the universe. Results and theorems, such as the properties of the integers and real numbers, do not depend in any way on the particular nature of reality in which we find ourselves. Mathematical truths would be true in any universe.

Yet both fields are similar. In physics, and indeed in science generally, scientists compress their experimental observations into scientific laws. They then show how their observations can be deduced from these laws. In mathematics, too, something like this happens— mathematicians compress their computational experiments into mathematical axioms, and they then show how to deduce theorems from these axioms.

If Hilbert had been right, mathematics would be a closed system, without room for new ideas. There would be a static, closed theory of everything for all of mathematics, and this would be like a dictatorship. In fact, for mathematics to progress you actually need new ideas and plenty of room for creativity. It does not suffice to grind away, mechanically deducing all the possible consequences of a fixed number of basic principles. I much prefer an open system. I do not like rigid, authoritarian ways of thinking.

Another person who thought mathematics is like physics was Imre Lakatos, who left Hungary in 1956 and later worked on philosophy of science in England. There Lakatos came up with a great word, “quasi-empirical,” which means that even though there are no true experiments that can be carried out in mathematics, something similar does take place. For example, the Goldbach conjecture states that any even number greater than 2 can be expressed as the sum of two prime numbers. This conjecture was arrived at experimentally, by noting empirically that it was true for every even number that anyone cared to examine. The conjecture has not yet been proved, but it has been verified up to 10^{14}.

I think that mathematics is quasi-empirical. In other words, I feel that mathematics is different from physics (which is truly empirical) but perhaps not as different as most people think.

I have lived in the worlds of both mathematics and physics, and I never thought there was such a big difference between these two fields. It is a matter of degree, of emphasis, not an absolute difference. After all,

mathematics and physics coevolved. Mathematicians should not isolate themselves. They should not cut themselves off from rich sources of new ideas.

New Mathematical Axioms

The idea of choosing to add more axioms is not an alien one to mathematics. A well-known example is the parallel postulate in Euclidean geometry: given a line and a point not on the line, there is exactly one line that can be drawn through the point that never intersects the original line. For centuries geometers wondered whether that result could be proved using the rest of Euclid's axioms. It could not. Finally, mathematicians realized that they could substitute different axioms in place of the Euclidean version, thereby producing the non-Euclidean geometries of curved spaces, such as the surface of a sphere or of a saddle.

Other examples are the law of the excluded middle in logic and the axiom of choice in set theory. Most mathematicians are happy to make use of those axioms in their proofs, although others do not, exploring instead so-called intuitionist logic or constructivist mathematics. Mathematics is not a single monolithic structure of absolute truth!

Another very interesting axiom may be the "P not equal to NP" conjecture. P and NP are names for classes of problems. An NP problem is one for which a proposed solution can be verified quickly. For example, for the problem "find the factors of 8,633," one can quickly verify the proposed solution "97 and 89" by multiplying those two numbers. (There is a technical definition of "quickly," but those details are not important here.) A P problem is one that can be solved quickly even without being given the solution. The question is—and no one knows the answer—can every NP problem be solved quickly? (Is there a quick way to find the factors of 8,633?) That is, is the class P the same as the class NP? This problem is one of the Clay Millennium Prize Problems for which a reward of $1 million is on offer.

Computer scientists widely believe that P is not equal to NP, but no proof is known. One could say that a lot of quasiempirical evidence points to P not being equal to NP. Should P not equal to NP be adopted as an axiom, then? In effect, this is what the computer science community has done.

Closely related to this issue is the security of certain cryptographic systems used throughout the world. The systems are believed to be invulnerable to being cracked, but no one can prove it.

Experimental Mathematics

Another area of similarity between mathematics and physics is experimental mathematics: the discovery of new mathematical results by looking at many examples using a computer. Whereas this approach is not as persuasive as a short proof, it can be more convincing than a long and extremely complicated proof, and for some purposes it is quite sufficient.

In the past, this approach was defended with great vigor by both George Pólya and Lakatos, believers in heuristic reasoning and in the quasi-empirical nature of mathematics. This methodology is also practiced and justified in Stephen Wolfram's *A New Kind of Science* (2002).

Extensive computer calculations can be extremely persuasive, but do they render proof unnecessary? Yes and no. In fact, they provide a different kind of evidence. In important situations, I would argue that both kinds of evidence are required, as proofs may be flawed, and conversely computer searches may have the bad luck to stop just before encountering a counterexample that disproves the conjectured result.

All these issues are intriguing but far from resolved. It is now 2006, 50 years after this magazine published its article on Gödel's proof, and we still do not know how serious incompleteness is. We do not know if incompleteness is telling us that mathematics should be done somewhat differently. Maybe 50 years from now we will know the answer.

--Originally published: Scientific American 294(3); 74-81 (March 2006).

The Unsolvable Problem

by Toby S. Cubitt, David Pérez-García and Michael Wolf

The three of us were sitting together in a café in Seefeld, a small town deep in the Austrian Alps. It was the summer of 2012, and we were stuck. Not stuck in the café—the sun was shining, the snow on the Alps was glistening, and the beautiful surroundings were sorely tempting us to abandon the mathematical problem we were stuck on and head outdoors. We were trying to explore the connections between 20th-century mathematical results by Kurt Gödel and Alan Turing and quantum physics. That, at least, was the dream. A dream that had begun back in 2010, during a semester-long program on quantum information at the Mittag-Leffler Institute near Stockholm.

Some of the questions we were looking into had been explored before by others, but to us this line of research was entirely new, so we were starting with something simple. Just then, we were trying to prove a small and not very significant result to get a feel for things. For months now, we had a proof (of sorts) of this result. But to make the proof work, we had to set up the problem in an artificial and unsatisfying way. It felt like changing the question to suit the answer, and we were not very happy with it. Picking the problem up again during the break after the first session of talks at the workshop in Seefeld that had brought us together in 2012, we still could not see any way around our problems. Half-jokingly, one of us (Michael Wolf) asked, “Why don’t we prove the undecidability of something people really care about, like the spectral gap?”

At the time, we were interested in whether certain problems in physics are “decidable” or “undecidable”—that is, can they ever be solved? We had gotten stuck trying to probe the decidability of a much more minor question, one few people care about. The “spectral gap” problem Michael was proposing that we tackle (which we will explain later) was one of

central importance to physics. We did not know at the time whether this problem was or was not decidable (although we had a hunch it was not) or whether we would be able to prove it either way. But if we could, the results would be of real relevance to physics, not to mention a substantial mathematical achievement. Michael's ambitious suggestion, tossed out almost as a jest, launched us on a grand adventure. Three years and 146 pages of mathematics later, our proof of the undecidability of the spectral gap was published in *Nature*.

To understand what this means, we need to go back to the beginning of the 20th century and trace some of the threads that gave rise to modern physics, mathematics and computer science. These disparate ideas all lead back to German mathematician David Hilbert, often regarded as the greatest figure of the past 100 years in the field. (Of course, no one outside of mathematics has heard of him. The discipline is not a good route to fame and celebrity, although it has its own rewards.)

The Mathematics of Quantum Mechanics

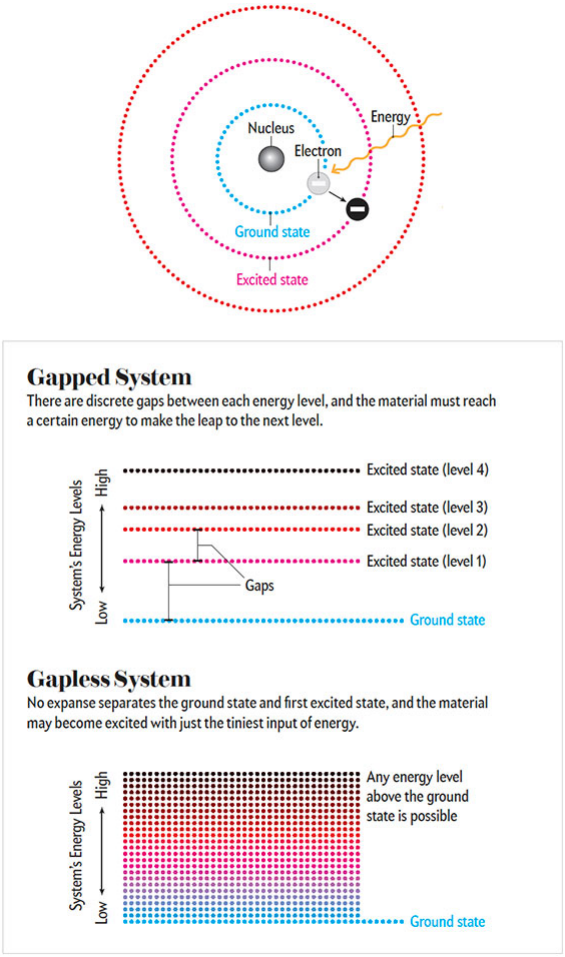
Hilbert's influence on mathematics was immense. Early on, he developed a branch of mathematics called functional analysis—in particular, an area known as spectral theory, which would end up being key to the question within our proof. Hilbert was interested in this area for purely abstract reasons. But as so often happens, his mathematics turned out to be exactly what was necessary to understand a question that was perplexing physicists at the time.

If you heat a substance up, it begins to glow as the atoms in it emit light (hence the phrase "red hot"). The yellow-orange light from sodium street lamps is a good example: sodium atoms predominantly emit light at a wavelength of 590 nanometers, in the yellow part of the visible spectrum. Atoms absorb or release light when electrons within them "jump" between energy levels, and the precise frequency of that light depends on the energy gap between the levels. The frequencies of light emitted by heated materials thus give us a "map" of the gaps between the atom's different energy levels. Explaining these atomic emissions was one of the problems physicists were wrestling with in the first half of the 20th century. The question led directly to the development of quantum mechanics, and the mathematics of Hilbert's spectral theory played a prime role.

One of these gaps between quantum energy levels is especially important. The lowest possible energy level of a material is called its ground state. This is the level it will sit in when it has no heat. To get a material into its ground state, scientists must cool it down to extremely low temperatures in a laboratory. Then, if the material is to do anything other than sit in its ground state, something must excite it to a higher energy. The easiest way is for it to absorb the smallest amount of energy it can, just enough to take it to the next energy level above the ground state—the first excited state. The energy gap between the ground state and this first excited state is so critical that it is often just called the spectral gap.

The Spectral Gap

The authors' mathematical proof took on the question of the "spectral gap"—the jump in energy between the ground state and first excited state of a material. When we think of energy states, we tend to think of electrons in atoms, which can jump up and down between energy levels. Whereas in atoms there is always a gap between such levels, in larger materials made of many atoms, there is sometimes no distance between the ground state and the first excited state: even the smallest possible amount of energy will be enough to push the material up an energy level. Such materials are called "gapless." The authors proved that it will never be possible to determine whether all materials are gapped or gapless.



Credit: Illustration by Jen Christiansen.

In some materials, there is a large gap between the ground state and the first excited state. In other materials, the energy levels extend all the way down to the ground state without any gaps at all. Whether a material is “gapped” or “gapless” has profound consequences for its behavior at low temperatures. It plays a particularly significant role in quantum phase transitions.

A phase transition happens when a material undergoes a sudden and dramatic change in its properties. We are all very familiar with some phase transitions—such as water transforming from its solid form of ice into its liquid form when heated up. But there are more exotic quantum phase transitions that happen even when the temperature is kept extremely low. For example, changing the magnetic field around a material or the pressure

it is subjected to can cause an insulator to become a superconductor or cause a solid to become a superfluid.

How can a material go through a phase transition at a temperature of absolute zero (–273.15 degrees Celsius), at which there is no heat at all to provide energy? It comes down to the spectral gap. When the spectral gap disappears—when a material is gapless—the energy needed to reach an excited state becomes zero. The tiniest amount of energy will be enough to push the material through a phase transition. In fact, thanks to the weird quantum effects that dominate physics at these very low temperatures, the material can temporarily "borrow" this energy from nowhere, go through a phase transition and "give" the energy back. Therefore, to understand quantum phase transitions and quantum phases, we need to determine when materials are gapped and when they are gapless.

Because this spectral gap problem is so fundamental to understanding quantum phases of matter, it crops up all over the place in theoretical physics. Many famous and long-standing open problems in condensed matter physics boil down to solving this problem for a specific material. A closely related question even crops up in particle physics: there is very good evidence that the fundamental equations describing quarks and their interactions have a "mass gap." Experimental data from particle colliders such as the Large Hadron Collider near Geneva support this notion, as do massive number-crunching results from supercomputers. But proving the idea rigorously from the theory seems to be extremely difficult. So difficult, in fact, that this problem, called the Yang-Mills mass gap problem, has been named one of seven Millennium Prize problems by the Clay Mathematics Institute, and anyone who solves it is entitled to a $1-million prize. All these problems are particular cases of the general spectral gap question. We have bad news for anyone trying to solve them, though. Our proof shows that the general problem is even trickier than we thought. The reason comes down to a question called the *Entscheidungsproblem*.

Unanswerable Questions

By the 1920s Hilbert had become concerned with putting the foundations of mathematics on a firm, rigorous footing—an endeavor that became known as Hilbert's program. He believed that whatever mathematical conjecture one might make, it will in principle be possible to prove either

that it is true or that it is false. (It had better not be possible to prove that it is both, or something has gone very wrong with mathematics!) This idea might seem obvious, but mathematics is about establishing concepts with absolute certainty. Hilbert wanted a rigorous proof.

In 1928 he formulated the *Entscheidungsproblem*. Although it sounds like the German sound for a sneeze, in English it translates to “the decision problem.” It asks whether there is a procedure, or “algorithm,” that can decide whether mathematical statements are true or false.

For example, the statement “Multiplying any whole number by 2 gives an even number” can easily be proved true, using basic logic and arithmetic. Other statements are less clear. What about the following example? “If you take any whole number and divide it by 2 if it’s even or multiply it by 3 and add 1 if it’s odd, and then repeat the process, you always eventually reach the number 1..” (Have a think about it.)

Unfortunately for Hilbert, his hopes were to be dashed. In 1931 Gödel published some remarkable results now known as his incompleteness theorems. Gödel showed that there are perfectly reasonable mathematical statements about whole numbers that can be neither proved nor disproved. In a sense, these statements are beyond the reach of logic and arithmetic. And he proved this assertion. If that is hard to wrap your head around, you are in good company. Gödel’s incompleteness theorems shook the foundations of mathematics to the core.

Here is a flavor of Gödel’s idea: If someone tells you, “This sentence is a lie,” is that person telling the truth or lying? If he or she is telling the truth, then the statement must indeed be a lie. But if he or she is lying, then it is true. This quandary is known as the liar paradox. Even though it appears to be a perfectly reasonable English sentence, there is no way to determine whether it is true or false. What Gödel managed to do was to construct a rigorous mathematical version of the liar paradox using only basic arithmetic.

The next major player in the story of the *Entscheidungsproblem* is Alan Turing, the English computer scientist. Turing is most famous among the general public for his role in breaking the German Enigma code during World War II. But among scientists, he is best known for his 1937 paper

“On Computable Numbers, with an Application to the *Entscheidungsproblem*.” Strongly influenced by Gödel’s result, the young Turing had given a negative answer to Hilbert’s *Entscheidungsproblem* by proving that no general algorithm to decide whether mathematical statements are true or false can exist. (American mathematician Alonzo Church also independently proved this just before Turing. But Turing’s proof was ultimately more significant. Often in mathematics, the proof of a result turns out to be more important than the result itself.)

To solve the *Entscheidungsproblem*, Turing had to pin down precisely what it meant to “compute” something. Nowadays we think of computers as electronic devices that sit on our desk, on our lap or even in our pocket. But computers as we know them did not exist in 1936. In fact, “computer” originally meant a person who carried out calculations with pen and paper. Nevertheless, computing with pen and paper as you did in high school is mathematically no different to computing with a modern desktop computer —just much slower and far more prone to mistakes.

Turing came up with an idealized, imaginary computer called a Turing machine. This very simple imaginary machine does not look like a modern computer, but it can compute everything that the most powerful modern computer can. In fact, any question that can ever be computed (even on quantum computers or computers from the 31st century that have yet to be invented) could also be computed on a Turing machine. It would just take the Turing machine much longer.

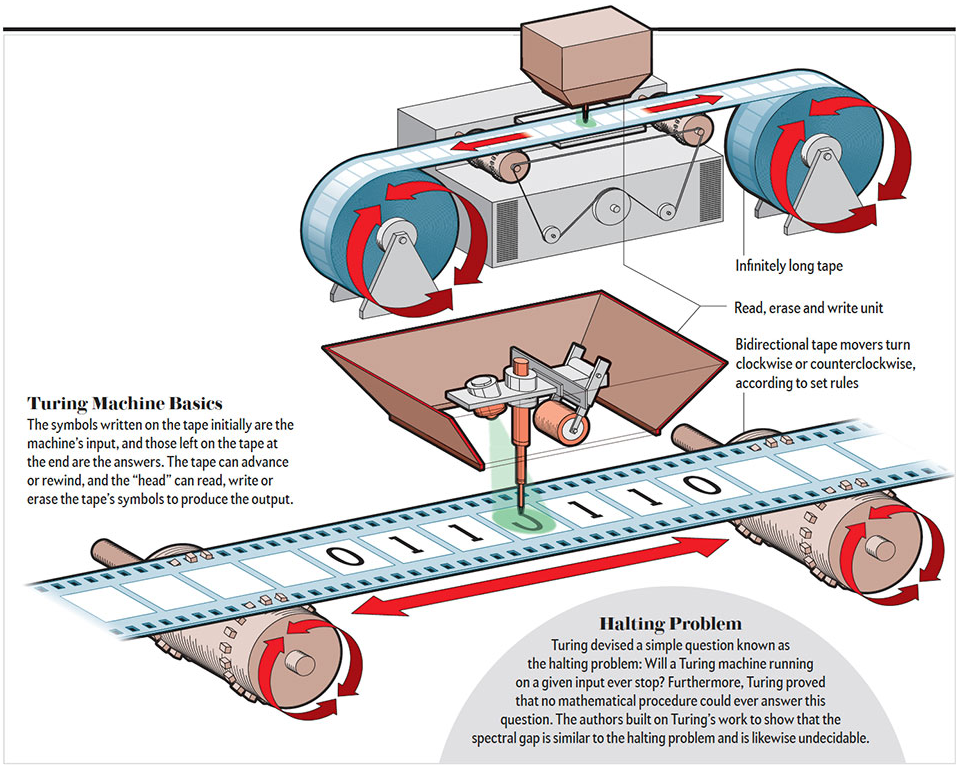
A Turing machine has an infinitely long ribbon of tape and a “head” that can read and write one symbol at a time on the tape, then move one step to the right or left along it. The input to the computation is whatever symbols are originally written on the tape, and the output is whatever is left written on it when the Turing machine finally stops running (halts). The invention of the Turing machine was more important even than the solution to the *Entscheidungsproblem*. By giving a precise, mathematically rigorous formulation of what it meant to make a computation, Turing founded the modern field of computer science.

Having constructed his imaginary mathematical model of a computer, Turing then went on to prove that there is a simple question about Turing machines that no mathematical procedure can ever decide: Will a Turing

machine running on a given input ever halt? This question is known as the halting problem. At the time, this result was shocking. Mathematicians have become accustomed to the fact that any conjecture we are working on could be provable, disprovable or undecidable.

Turing Machine

Before modern computers existed, mathematician Alan Turing imagined a hypothetical device called a Turing machine that defined what it meant to "compute." The machine reads and performs operations on the symbols written on an infinitely long strip of tape that runs through it. The concept turned out to be central to the authors' proof of the undecidability of the spectral gap problem.



Credit: Illustration by Ben Gilliland

Where We Come In

In our result, we had to tie all these disparate threads back together. We wanted to unite the quantum mechanics of the spectral gap, the computer science of undecidability and Hilbert's spectral theory to prove that—like the halting problem—the spectral gap problem was one of the undecidable ones that Gödel and Turing taught us about.

Chatting in that café in Seefeld in 2012, we had an idea for how we might be able to prove a weaker mathematical result related to the spectral gap. We tossed this idea around, not even scribbling on the back of a napkin, and it seemed like it might work. Then the next session of talks started. And there we left it.

A few months later one of us (Toby Cubitt) visited Michael in Munich, and we did what we had not done in Seefeld: jotted some equations down on a scrap of paper and convinced ourselves the idea worked. In the following weeks, we completed the argument and wrote it up properly in a private four-page note. (Nothing in mathematics is truly proved until you write it down—or, better still, type it up and show it to a colleague for scrutiny.) Conceptually this was a major advance. Before now, the idea of proving the undecidability of the spectral gap was more of a joke than a serious prospect. Now we had the first glimmerings that it might actually be possible. But there was still a very long way to go. We could not extend our initial idea to prove the undecidability of the spectral gap problem itself.

Burning the Midnight Coffee

We attempted to make the next leap by linking the spectral gap problem to quantum computing. In 1985 Nobel Prize–winning physicist Richard Feynman published one of the papers that launched the idea of quantum computers. In that paper, Feynman showed how to relate ground states of quantum systems to computation. Computation is a dynamic process: you supply the computer with input, and it goes through several steps to compute a result and outputs the answer. But ground states of quantum systems are completely static: the ground state is just the configuration a material sits in at zero temperature, doing nothing at all. So how can it make a computation?

The answer comes through one of the defining features of quantum mechanics: superposition, which is the ability of objects to occupy many states simultaneously, as, for instance, Erwin Schrödinger’s famous quantum cat can be alive and dead at the same time. Feynman proposed constructing a quantum state that is in a superposition of the various steps in a computation— initial input, every intermediate step of the computation and final output—all at once. Alexei Kitaev of the California Institute of

Technology later developed this idea substantially by constructing an imaginary quantum material whose ground state looks exactly like this.

If we used Kitaev's construction to put the entire history of a Turing machine into the material's ground state in superposition, could we transform the halting problem into the spectral gap problem? In other words, could we show that any method for solving the spectral gap problem would also solve the halting problem? Because Turing had already shown that the halting problem was undecidable, this would prove that the spectral gap problem must also be undecidable.

Encoding the halting problem in a quantum state was not a new idea. Seth Lloyd, now at the Massachusetts Institute of Technology, had proposed this almost two decades earlier to show the undecidability of another quantum question. Daniel Gottesman of the Perimeter Institute for Theoretical Physics in Waterloo and Sandy Irani of the University of California, Irvine, had used Kitaev's idea to prove that even single lines of interacting quantum particles can show very complex behavior. In fact, it was Gottesman and Irani's version of Kitaev's construction that we hoped to make use of.

But the spectral gap is a different kind of problem, and we faced some apparently insurmountable mathematical obstacles. The first had to do with supplying the input into the Turing machine. Remember that the undecidability of the halting problem is about whether the Turing machine halts *on a given input*. How could we design our imaginary quantum material in a way that would let us choose the input to the Turing machine to be encoded in the ground state?

When working on that earlier problem (the one we were still stuck on in the café in Seefeld), we had an idea of how to rectify the issue by putting a "twist" in the interactions between the particles and using the angle of this rotation to create an input to the Turing machine. In January 2013 we met at a conference in Beijing and discussed this plan together. But we quickly realized that what we had to prove came very close to contradicting known results about quantum Turing machines. We decided we needed a complete and rigorous proof that our idea worked before we pursued the project further.

At this point, Toby had been part of David Pérez-García's group at Complutense University of Madrid for more than two years. In that same month he moved to the University of Cambridge, but his new apartment there was not yet ready, so his friend and fellow quantum information theorist Ashley Montanaro offered to put him up. For those two months, he set to work producing a rigorous proof of this idea. His friend would find him at the kitchen table in the morning, a row of empty coffee mugs next to him, about to head to bed, having worked through the night figuring out details and typing them up. At the end of those two months, Toby sent around the completed proof.

In Rememberance of Tilings Past

This 29-proof showed how to overcome one of the obstacles to connecting the ground state of a quantum material to computation with a Turing machine. But there was an even bigger obstacle to that goal: the resulting quantum material was always gapless. If it is always gapless, the spectral gap problem for this particular material is very easy to solve: the answer is gapless!

Our first idea from Seefeld, which proved a much weaker result than we wanted, nonetheless managed to get around this obstacle. The key was using "tilings." Imagine you are covering a large bathroom floor with tiles. In fact, imagine it is an infinitely big bathroom. The tiles have a very simple pattern on them: each of the four sides of the tile is a different color. You have various boxes of tiles, each with a different arrangement of colors. Now imagine there is an infinite supply of tiles in each box. You, of course, want to tile the infinite bathroom floor so that the colors on adjacent tiles match. Is this possible?

The answer depends on which boxes of tiles you have available. With some sets of colored tiles, you will be able to tile the infinite bathroom floor. With others, you will not. Before you select which boxes of tiles to buy, you would like to know whether or not they will work. Unfortunately for you, in 1966 mathematician Robert Berger proved that this problem is undecidable.

One easy way to tile the infinite bathroom floor would be to first tile a small rectangle so that colors on opposite sides of it match. You could then

cover the entire floor by repeating this rectangular pattern. Because they repeat every few tiles, such patterns are called periodic. The reason the tiling problem is undecidable is that nonperiodic tilings also exist: patterns that cover the infinite floor but never repeat.

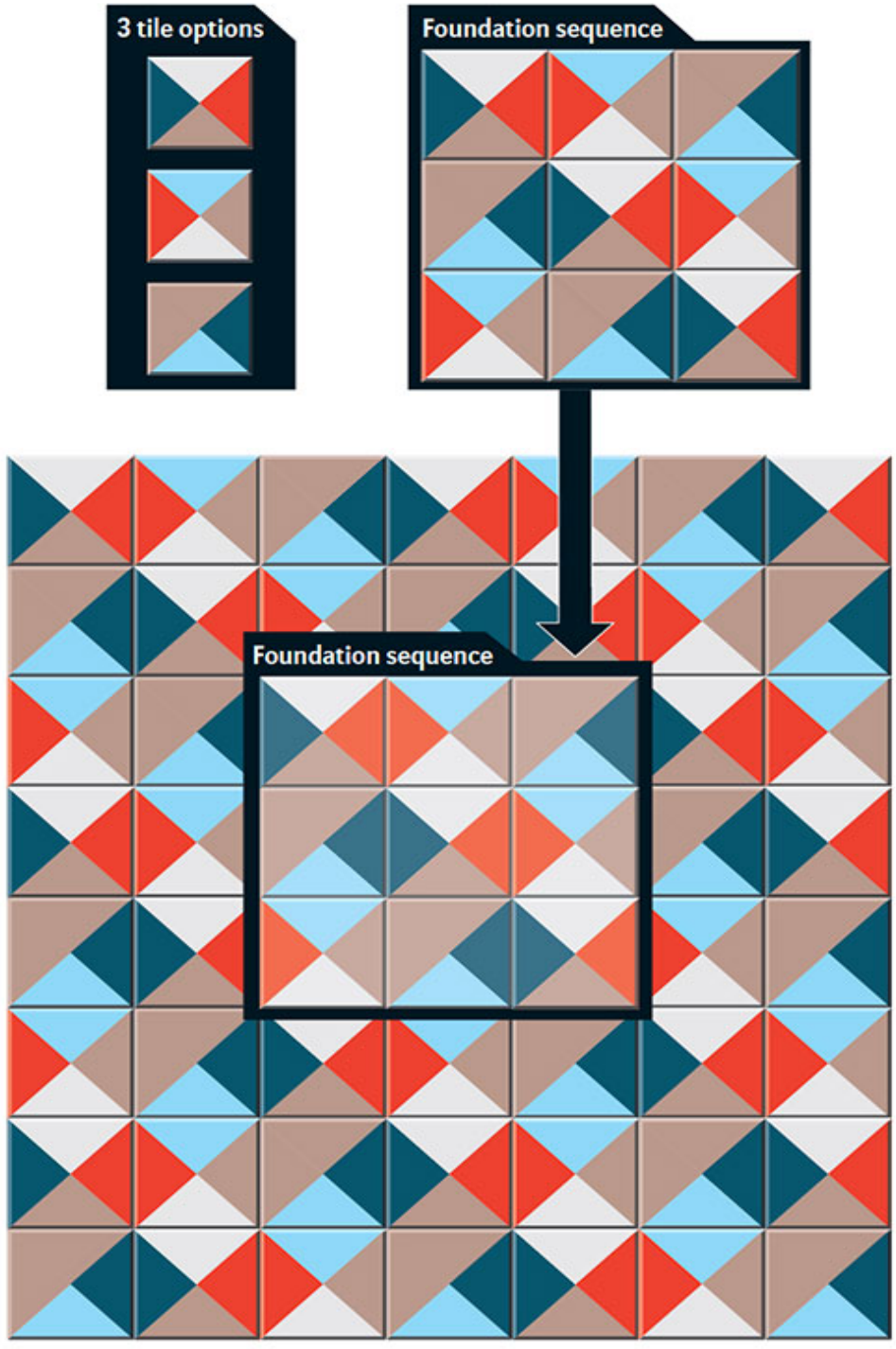
Back when we were discussing our first small result, we studied a 1971 simplification of Berger's original proof made by Rafael Robinson of the University of California, Berkeley. Robinson constructed a set of 56 different boxes of tiles that, when used to tile the floor, produce an interlocking pattern of ever larger squares. This fractal pattern looks periodic, but in fact, it never quite repeats itself. We extensively discussed ways of using tiling results to prove the undecidability of quantum properties. But back then, we were not even thinking about the spectral gap. The idea lay dormant.

Tiling an Infinite Bathroom Floor

To connect the spectral gap problem to the halting problem, the authors considered the classic mathematical question of how to tile an infinitely large floor. Imagine you have a box with a certain selection of tiles, and you want to arrange them so that the colors on the sides of each tile match those next to them. In some cases, this is possible by tiling the floor in either a repeating "periodic" pattern or a fractallike "aperiodic" pattern.

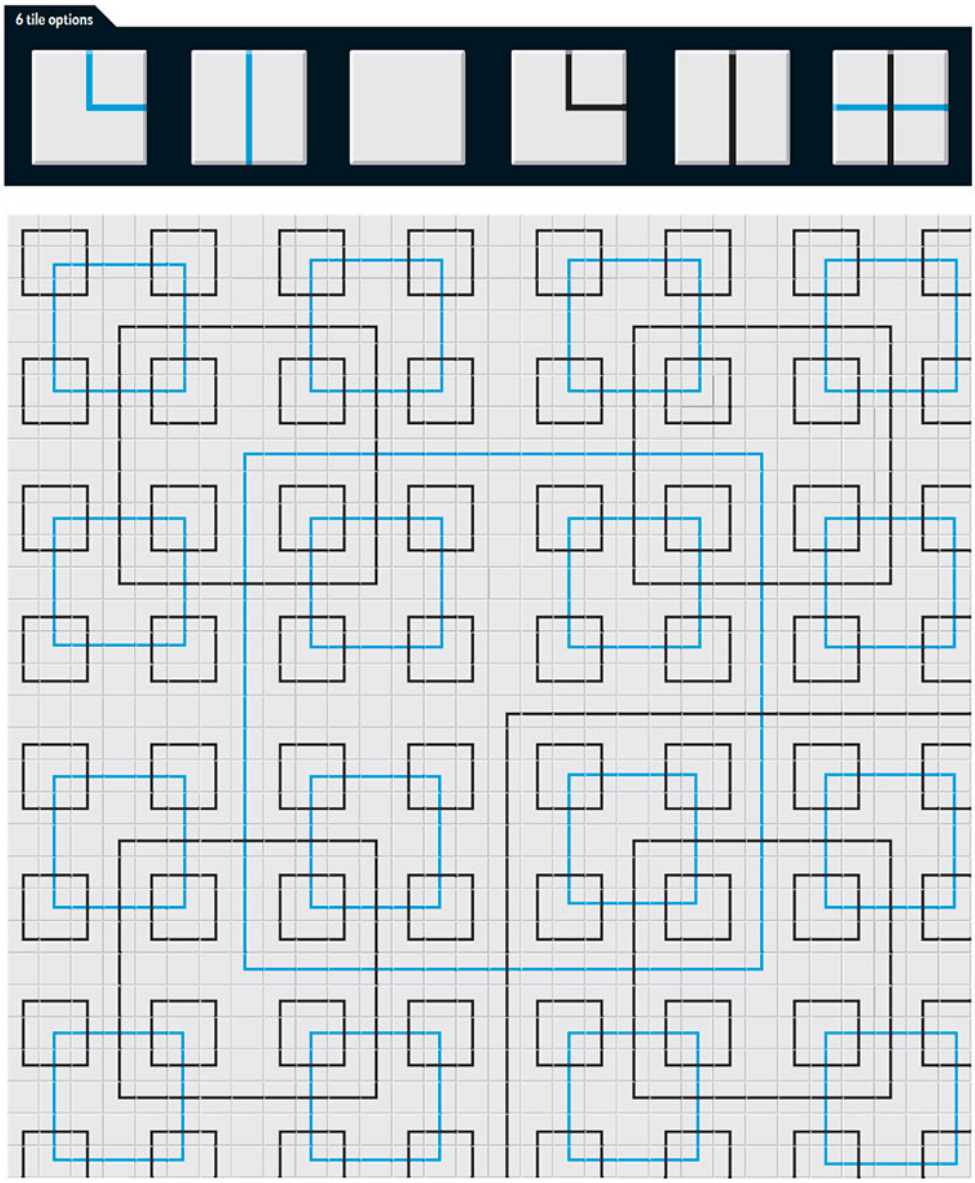
Periodic Tiles

One version of the classic problem concerns tiles that come in three varieties containing five different colors. In this particular case, it is possible to tile the floor with all sides matching up by creating a rectangle that repeats. On each side of the rectangle, the colors match so that many versions of the same rectangle can be placed next to one another in an infinite pattern.



Aperiodic Tiles

In their proof, the authors used a particular set of tiles designed by mathematician Rafael Robinson in 1971. Robinson's tiles fit together in an ever expanding sequence that does not quite repeat but instead creates a fractal-like pattern. All rotations of the six tiles shown here are allowed. There are also other ways to fit these pieces together in a periodic pattern, but by adding more markings to these tiles (*not shown*), Robinson designed a set of 56 tiles for which no pattern is possible other than the one shown.



Credit: Illustration by Jen Christiansen.

In April 2013 Toby paid a visit to Charlie Bennett at IBM's Thomas J. Watson Research Center. Among Bennett's many achievements before becoming one of the founding fathers of quantum information theory was his seminal 1970s work on Turing machines. We wanted to quiz him about some technical details of our proof to make sure we were not overlooking something. He said he had not thought about this stuff for 40 years, and it was high time a younger generation took over. (He then went on to very

helpfully explain some subtle mathematical details of his 1970s work, which reassured us that our proof was okay.)

Bennett has an immense store of scientific knowledge. Because we had been talking about Turing machines and undecidability, he e-mailed copies of a couple of old papers on undecidability he thought might interest us. One of these was the same 1971 paper by Robinson that we had studied. Now the time was right for the ideas sowed in our earlier discussions to spring to life. Reading Robinson's paper again, we realized it was exactly what we needed to prevent the spectral gap from vanishing.

Our initial idea had been to encode one copy of the Turing machine into the ground state. By carefully designing the interactions between the particles, we could make the ground state energy a bit higher if the Turing machine halted. The spectral gap—the energy jump to the first excited state—would then depend on whether the Turing machine halted or not. There was just one problem with this idea, and it was a big one. As the number of particles increased, the additional contribution to the ground state energy got closer and closer to zero, leading to a material that was always gapless.

But by adapting Berger's tiling construction, we could instead encode *many copies* of exactly the same Turing machine into the ground state. In fact, we could attach one copy to each square in Robinson's tiling pattern. Because these are identical copies of the same Turing machine, if one of them halts, they all halt. The energy contributions from all these copies add up. As the number of particles increases, the number of squares in the tiling pattern gets bigger. Thus, the number of copies of the Turing machine increases, and their energy contribution becomes huge, giving us the possibility of a spectral gap.

Exams and Deadlines

One significant weakness remained in the result we had proved. We could not say anything about how big the energy gap was when the material was gapped. This uncertainty left our result open to the criticism that the gap could be so small that it might as well not exist. We needed to prove that the gap, when it existed, was actually large. The first solution we found arose by considering materials in three dimensions instead of the planar materials we had been thinking about until then.

When you cannot stop thinking about a mathematical problem, you make progress in the most unexpected places. David worked on the details of this idea in his head while he was supervising an exam. Walking along the rows of tables in the hall, he was totally oblivious to the students working feverishly around him. Once the test was over, he committed this part of the proof to paper.

We now knew that getting a big spectral gap was possible. Could we also get it in two dimensions, or were three necessary? Remember the problem of tiling an infinite bathroom floor. What we needed to show was that for the Robinson tiling, if you got one tile wrong somewhere, but the colors still matched everywhere else, then the pattern formed by the tiles would be disrupted only in a small region centered on that wrong tile. If we could show this “robustness” of the Robinson tiling, it would imply that there was no way of getting a small spectral gap by breaking the tiling only a tiny bit.

By the late summer of 2013, we felt we had all the ingredients for our proof to work. But there were still some big details to be resolved, such as proving that the tiling robustness could be merged with all the other proof ingredients to give the complete result. The Isaac Newton Institute for Mathematical Science in Cambridge, England, was hosting a special workshop on quantum information for the whole of the autumn semester of 2013. All three of us were invited to attend. It was the perfect opportunity to work together on finishing the project. But David was not able to stay in Cambridge for long. We were determined to complete the proof before he left.

The Isaac Newton Institute has blackboards everywhere—even in the bathrooms! We chose one of the blackboards in a corridor (the closest to the coffee machine) for our discussions. We spent long hours at the blackboard developing the missing ideas, then divided the task of making these ideas mathematically rigorous among us. This process always takes far more time and effort than it seems on the blackboard. As the date of David’s departure loomed, we worked without interruption all day and most of the night. Just a few hours before he left for home, we finally had a complete proof.

In physics and mathematics, researchers make most results public for the first time by posting a draft paper to the arXiv.org preprint server before submitting it to a journal for peer review. Although we were now fairly

confident the entire argument worked and the hardest part was behind us, our proof was not ready to be posted. There were many mathematical details to be filled in. We also wanted to rewrite and tidy up the paper (we hoped to reduce the page count in the process, although in this we would completely fail). Most important, although at least one of us had checked every part of the proof, no one had gone through it all from beginning to end.

In summer 2014 David was on a sabbatical at the Technical University of Munich with Michael. Toby went out to join them. The plan was to spend this time checking and completing the whole proof, line by line. David and Toby were sharing an office. Each morning David would arrive with a new printout of the draft paper, copious notes and questions scribbled in the margins and on interleaved sheets. The three of us would get coffee and then pick up where we had left off the day before, discussing the next section of the proof at the blackboard. In the afternoon, we divided up the work of rewriting the paper and adding the new material and of going through the next section of the proof. Toby was suffering from a slipped disc and could not sit down, so he worked with his laptop propped on top of an upturned garbage bin on top of the desk. David sat opposite, the growing pile of printouts and notes taking up more and more of his desk. On a couple of occasions, we found significant gaps in the proof. These turned out to be surmountable, but bridging them meant adding substantial material to it. The page count continued to grow.

After six weeks, we had checked, completed and improved every single line of the proof. It would take another six months to finish writing everything up. Finally, in February 2015, we uploaded the paper to arXiv.org.

What It All Means

Ultimately what do these 146 pages of complicated mathematics tell us?

First, and most important, they give a rigorous mathematical proof that one of the basic questions of quantum physics cannot be solved in general. Note that the “in general” here is critical. Even though the halting problem is undecidable in general, for *particular* inputs to a Turing machine, it is often still possible to say whether it will halt or not. For example, if the first

instruction of the input is “halt,” the answer is pretty clear. The same goes if the first instruction tells the Turing machine to loop forever. Thus, although undecidability implies that the spectral gap problem cannot be solved for *all* materials, it is entirely possible to solve it for specific materials. In fact, condensed matter physics is littered with such examples. Nevertheless, our result proves rigorously that even a perfect, complete description of the microscopic interactions between a material’s particles is not always enough to deduce its macroscopic properties.

You may be asking yourself if this finding has any implications for “real physics.” After all, scientists can always try to measure the spectral gap in experiments. Imagine if we could engineer the quantum material from our mathematical proof and produce a piece of it in the lab. Its interactions are so extraordinarily complicated that this task is far, far beyond anything scientists are ever likely to be able to do. But if we could and then took a piece of it and tried to measure its spectral gap, the material could not simply throw up its hands and say, “I can’t tell you—it’s undecidable.” The experiment would have to measure *something*.

The answer to this apparent paradox lies in the fact that, strictly speaking, the terms “gapped” and “gapless” only make mathematical sense when the piece of material is infinitely large. Now, the 10^{23} or so atoms contained in even a very small piece of material represent a very large number indeed. For normal materials, this is close enough to infinity to make no difference. But for the very strange material constructed in our proof, large is not equivalent to infinite. Perhaps with 10^{23} atoms, the material appears in experiments to be gapless. Just to be sure, you take a sample of material twice the size and measure again. Still gapless. Then, late one night, your graduate student comes into the lab and adds just one extra atom. The next morning, when you measure it again, the material has become gapped! Our result proves that the size at which this transition may occur is incomputable (in the same Gödel-Turing sense that you are now familiar with). This story is completely hypothetical for now because we cannot engineer a material this complex. But it shows, backed by a rigorous mathematical proof, that scientists must take special care when extrapolating experimental results to infer the behavior of the same material at larger sizes.

And now we come back to the Yang-Mills problem—the question of whether the equations describing quarks and their interactions have a mass gap. Computer simulations indicate that the answer is yes, but our result suggests that determining for sure may be another matter. Could it be that the computer-simulation evidence for the Yang-Mills mass gap would vanish if we made the simulation just a tiny bit larger? Our result cannot say, but it does open the door to the intriguing possibility that the Yang-Mills problem, and other problems important to physicists, may be undecidable.

And what of that original small and not very significant result we were trying to prove all those years ago in a café in the Austrian Alps? Actually, we are still working on it.

--Originally published: Scientific American 319(4); 28-37 (October 2018).